1 STABILITY, PERFORMANCE EVALUATION, AND OPTIMIZATION

Sean P. Meyn

Department of Electrical and Computer Engineering
and the Coordinated Sciences Laboratory
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Abstract: The theme of this chapter is stability and performance approximation for MDPs on an infinite state space. The main results are centered around stochastic Lyapunov functions for verifying stability and bounding performance. An operator-theoretic framework is used to reduce the analytic arguments to the level of the finite state-space case.

1.1 INTRODUCTION

1.1.1 Models on a general state space

This chapter focuses on stability of Markov chain models. Our main interest is the various relationships between stability; the existence of Lyapunov functions; performance evaluation; and existence of optimal policies for controlled Markov chains. We also consider two classes of algorithms for constructing policies, the policy and value iteration algorithms, since they provide excellent examples of the application of Lyapunov function techniques for ψ-irreducible Markov chains on an uncountable state space.

Considering the importance of these topics, it is not surprising that considerable research has been done in each of these directions. In this chapter we do not attempt a survey of all existing literature, or present the most comprehen-

sive results. In particular, only the average cost optimality criterion is treated, and the assumptions we impose imply that the average cost is independent of the starting point of the process. By restricting attention in this way we hope that we can make the methodology more transparent.

One sees in several chapters in this volume that the generalization from finite state spaces to countable state spaces can lead to considerable technicalities. In particular, invariant distributions may not exist, and the cost functions of interest may not take on finite values. It would be reasonable to assume that the move from countable state spaces, to MDPs on a general state space should be at least as difficult. This assumption is probably valid if one desires a completely general theory.

However, the MDPs that we typically come across in practice exhibit structure which simplifies analysis, sometimes bringing us to the level of difficulty found in the countable, or even the finite state space case. For example, all of the specific models to be considered in this chapter, and most in this volume, have some degree of spatial homogeneity. The processes found in most applications will also exhibit some level of continuity in the sense that from similar starting points, and similar control sequences, two realizations of the model will have similar statistical descriptions. We do not require strong continuity conditions such as the strong Feller property, although this assumption is sometimes useful to establish existence and uniqueness of solutions to the various static optimization problems that arise in the analysis of controlled Markov chains. An assumption of ψ-irreducibility, to be described and developed below, allows one to lift much of the stability theory in the discrete state space setting to models on a completely general, non-countable state space. This is an exceptionally mild assumption on the model and, without this assumption, the theory of MDPs on a general state space is currently extremely weak.

1.1.2 An operator-theoretic framework

When $\boldsymbol{x}$ is ψ-irreducible it is possible to enlarge the state space to construct an *atom* $\boldsymbol{\theta} \in \mathbb{X}$ which is reachable from any initial condition (i.e. $\mathbb{P}_x\{\tau_\theta < \infty\} > 0$, $x \in \mathbb{X}$). When the atom is *recurrent*, that is, $\mathbb{P}_\theta\{\tau_\theta < \infty\} = 1$, then an invariant measure (see (1.9)) is given by

$$\mu\{Y\} = \mathbb{E}_\theta\Big[\sum_{t=1}^{\tau_\theta} \mathbf{1}_Y(x_t)\Big], \qquad Y \in \mathbb{F}, \tag{1.1}$$

where τ_θ is the first return time to $\boldsymbol{\theta}$ (see (1.6)), and $\mathbf{1}_Y$ is the indicator function of the set Y. This construction, and related results may be found in [45, 39]. In words, the quantity $\mu\{Y\}$ expresses the mean number of times that the chain visits the set Y before returning to $\boldsymbol{\theta}$. This expression assumes that τ_θ is almost surely finite. If the *mean* return time $\mathbb{E}_\theta[\tau_\theta]$ is finite then in fact the measure μ is finite, and it can then be normalized to give an invariant probability measure. Finiteness of the mean return time to some desirable state is the standard stability condition used for Markov chains, and for MDPs in which one is interested in the average cost optimality criterion.

Unfortunately, the split chain construction is cumbersome when developing a theory for controlled Markov chains. The sample path interpretation given in (1.1) is appealing, but it will be more convenient to work within an operator-theoretic framework, following [45]. To motivate this, suppose first that we remain in the previous setting with an uncontrolled Markov chain, and suppose that do have an state $\boldsymbol{\theta} \in \mathbb{X}$ satisfying $\mathbb{P}_\theta\{\tau_\theta < \infty\} = 1$. Denote by s the function which is equal to one at $\boldsymbol{\theta}$, and zero elsewhere: That is, $s = \mathbf{1}_\theta$. We let ν denote the probability measure on $\mathbb{X}$ given by $\nu(Y) = p(Y \mid \boldsymbol{\theta})$, $Y \in \mathbb{F}$, and define the 'outer product' of s and ν by

$$s \otimes \nu\,(x, Y) := s(x)\nu(Y).$$

For example, in the finite state space case the measure ν can be interpreted as a row vector, the function s as a column vector, and $s \otimes \nu$ is the standard (outer) product of these two vectors. Hence $s \otimes \nu$ is an $N \times N$ matrix, where N is the number of states.

In general, the *kernel* $s \otimes \nu$ may be viewed as a rank-one operator which maps L_∞ to itself, where L_∞ is the set of bounded, measurable functions on $\mathbb{X}$. Several other bounded linear operators on L_∞ will be developed in this chapter. The most basic are the n-step transition kernels, defined for $n \geq 1$ by

$$P^n f\,(x) := \int_{\mathbb{X}} f(y) p^n(dy|x), \qquad f \in L_\infty,$$

where $p^n(\cdot \mid \cdot)$ is the n-step state transition function for the chain. We set $P = P^1$. We can then write, in operator-theoretic notation,

$$\mathbb{P}_\theta\{\tau_\theta \geq n, x_n \in Y\} = \nu(P - s \otimes \nu)^{n-1}\mathbf{1}_Y, \qquad n \geq 1,$$

and hence the invariant measure μ given in (1.1) is expressed in this notation as

$$\mu(Y) = \sum_{n=1}^{\infty} \nu(P - s \otimes \nu)^{n-1}\mathbf{1}_Y, \qquad Y \in \mathbb{F}. \tag{1.2}$$

It is this algebraic description of μ that will be generalized and exploited in this chapter.

How can we mimic this algebraic structure without constructing an atom $\boldsymbol{\theta}$? First, we require a function $s\colon \mathbb{X} \to \mathbb{R}_+$ and a probability measure ν on $\mathbb{F}$ satisfying the *minorization condition*,

$$p(Y \mid x) \geq s(x)\nu(Y), \qquad x \in \mathbb{X},\ Y \in \mathbb{F}.$$

In operator theoretic notation this is written $P \geq s \otimes \nu$, and in the countable state space case this means that the transition matrix P dominates an outer product of two vectors with non-negative entries.

Unfortunately, this 'one step' minorization assumption excludes a large class of models, even the simple linear models to be considered as examples below. One can however move to the resolvent kernel defined by

$$K = (1 - \beta)\sum_{t=0}^{\infty} \beta^t P^t, \tag{1.3}$$

where $\beta \in]0,1[$ is some fixed constant. For a ψ-irreducible chain the required minorization always holds for the resolvent K [39, Theorem 5.2.3]. The move to the resolvent is useful since almost any object of interest can be mapped between the resolvent chain, and the original Markov chain. In particular, the invariant measures for P and K coincide (see [39, Theorem 10.4.3], or consider the resolvent equation in (1.11) below).

Much of the analysis then will involve the *potential kernels*, defined via

$$G := \sum_{t=0}^{\infty} K^t. \tag{1.4}$$

$$H := \sum_{t=0}^{\infty} (K - s \otimes \nu)^t. \tag{1.5}$$

In Theorem 1.1 below we demonstrate invariance of the σ-finite measure μ defined by,

$$\mu(Y) = \int_{\mathbb{X}} \nu(dx) H(x, Y), \qquad Y \in \mathbb{F},$$

provided the chain is *recurrent*. The invariant measure can be written in the compact form $\mu = \nu H$.

The measure μ will be *finite*, rather than just σ-finite, provided appropriate stability conditions are satisfied. The most natural stability assumption is equivalent to the existence of a Lyapunov function, whose form is very similar to the Poisson equation found in the average cost optimality equation. The development of these connections is one of the main themes of this chapter.

1.1.3 Overview

We conclude with an outline of the topics to follow. In the next section we review a bit of the general theory of ψ-irreducible chains, and develop some stochastic Lyapunov theory for such chains following [39, Chapters 11-14]. Following this, in Section 1.3 we develop in some detail the computation of the average cost through the Poisson equation, and the construction of bounds on the average cost. All of these results are developed for time homogeneous chains without control.

In Section 1.4 this stability theory is applied to the analysis of the average cost optimality equation (ACOE). We explore the consequences of this equation, and derive criteria for the existence of a solution.

Section 1.5 concerns two recursive algorithms for generating solutions to the ACOE: value iteration and policy iteration. It is shown that (i) either algorithm generates stabilizing stationary policies; (ii) for any of these policies, the algorithms generate uniform bounds on steady state performance. However, such results hold only if the algorithms are properly initialized.

Convergence is established for the policy iteration algorithm: Under suitable conditions, and when properly initialized, the algorithm converges to a solution of the ACOE.

Section 1.6 illustrates the theory with a detailed application to linear models, and to network scheduling.

This chapter is concluded with a discussion of some extensions and open problems.

1.2 STABILITY

In this section we consider a Markov chain $\boldsymbol{x}$ with uncontrolled transition function $p(\cdot \mid \cdot)$. The state space $\mathbb{X}$ is assumed to be a locally compact, separable metric space, and we let $\mathbb{F}$ denote the (countably generated) Borel σ-field on $\mathbb{X}$. Unless other references are given, all of the results described here together with their derivations can be found in [39].

1.2.1 ψ-irreducibility

Throughout this chapter we assume that ψ is a σ-finite measure on $\mathbb{F}$.

Definition 1.1

(i) *The chain is called* ψ***-irreducible*** *if the resolvent kernel defined in (1.3) satisfies*

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

We then call ψ an ***irreducibility measure.***

(ii) *We let $\mathbb{F}^+$ denote the set of all measurable $h \colon \mathbb{X} \to \mathbb{R}_+$ satisfying*

$$\psi(h) := \int_{\mathbb{X}} h(x)\psi(dx) > 0.$$

For $Y \in \mathbb{F}$ we write $Y \in \mathbb{F}^+$ provided $\psi(Y) > 0$.

If the chain is ψ-irreducible, then from any initial condition x, the process has a chance of entering any set in $\mathbb{F}^+$ in the sense that $\mathbb{P}_x\{\tau_Y < \infty\} > 0$, where τ_Y is the first return time,

$$\tau_Y = \min\{t \geq 1 : x_t \in Y\}. \tag{1.6}$$

Definition 1.2

(i) *A function $s \colon \mathbb{X} \to \mathbb{R}_+$ and a probability measure ν on $\mathbb{F}$ are called* ***petite*** *if*

$$K(x, Y) \geq s(x)\nu(Y), \qquad x \in \mathbb{X},\ Y \in \mathbb{F}. \tag{1.7}$$

(ii) *A set $Z \in \mathbb{F}$ is called* ***petite*** *if for some probability measure ν, and a constant $\delta > 0$,*

$$K(x, Y) \geq \delta\nu(Y), \qquad x \in Z,\ Y \in \mathbb{F}.$$

(iii) *The Markov chain is called a* T***-chain*** *if every compact set is petite.*

It is not difficult to show that, for a ψ-irreducible chain, the set Z is petite if for each $Y \in \mathbb{F}^+$, there exists $n \geq 1$, and $\delta > 0$ such that

$$\mathbb{P}_x(\tau_Y \leq n) \geq \delta \qquad \text{for any } x \in Z. \tag{1.8}$$

For a ψ-irreducible chain, there always exists a countable covering of the state space by petite sets. In virtually all examples these can be taken to be compact, so that $\boldsymbol{x}$ is a T-chain. The following result is taken from [39, Proposition 5.5.5]:

Proposition 1.1 *Suppose that the Markov chain $\boldsymbol{x}$ is ψ-irreducible. Then there is a non-negative function s and a probability measure ν satisfying (1.7). We can choose s so that it is strictly-positive valued, $s\colon \mathbb{X} \to]0,1[$, and ν can be chosen so that it is equivalent to ψ (that is, $\psi(A) = 0 \Leftrightarrow \nu(A) = 0$).* ∎

The bound (1.7) is the most powerful consequence of the ψ-irreducibility assumption since it allows the construction of the potential kernel H defined in (1.5). The following lemma will be useful below when constructing solutions to dynamic programming equations:

Lemma 1.1 *For any petite pair s, ν we have,*

$$Hs\,(x) := \sum_{i=0}^{\infty} (K - s \otimes \nu)^i s\,(x) \leq 1, \qquad x \in \mathbb{X}.$$

Proof. Define for $n \geq 0$, the kernel

$$H_n := \sum_{i=0}^{n} (K - s \otimes \nu)^i.$$

We show by induction that $H_n s$ is uniformly bounded for each n. For $n = 0$ we have $H_n s = s$, which is assumed to be bounded by one.

Suppose that $\|H_n s\|_\infty \leq 1$ for some arbitrary $n \geq 0$. We then have,

$$\begin{aligned} H_{n+1} s &= s + (K - s \otimes \nu) H_n s \\ &\leq s + (K - s \otimes \nu)\mathbf{1} \\ &= s + K\mathbf{1} - s \cdot \nu(\mathbb{X}) = K\mathbf{1} = 1. \end{aligned}$$

This proves the result since $H_n s \to Hs$ as $n \to \infty$. ∎

1.2.2 Recurrence

The crudest form of stability for a Markov chain is the property that the state visit 'important' sets with probability one from any starting point.

Definition 1.3

(i) *A ψ-irreducible chain is called* ***recurrent*** *if*

$$\mathbb{E}_x\Big[\sum_{t=0}^{\infty} \mathbf{1}(x_t \in Y)\Big] = \infty, \qquad Y \in \mathbb{F}^+,\ x \in \mathbb{X}.$$

(ii) *A measure μ on $\mathbb{F}$ is* ***invariant*** *if*

$$\mu(Y) = \int p(Y \mid x)\mu(dx), \qquad Y \in \mathbb{F}. \tag{1.9}$$

(iii) *If $\boldsymbol{x}$ is recurrent, and if in addition the chain admits an invariant probability measure μ, then the chain is called* ***positive recurrent.***

In terms of the potential kernel (1.4), recurrence is expressed,

$$G(x, Y) = \infty, \qquad Y \in \mathbb{F}^+,\ x \in \mathbb{X}.$$

There are several equivalent characterizations of recurrence which are easier to verify. A proof of the following equivalences can be found in [39, Chapter 8], or [45, Theorem 3.7].

Theorem 1.1 *The following are equivalent for a ψ-irreducible Markov chain $\boldsymbol{x}$:*

(i) *$\boldsymbol{x}$ is recurrent.*

(ii) *There exists a set $\mathbb{X}_0 \in \mathbb{F}$ satisfying $\psi\{\mathbb{X}_0^c\} = 0$, and*

$$\mathbb{P}_x\{\tau_Y < \infty\} = 1, \qquad Y \in \mathbb{F}^+,\ x \in \mathbb{X}_0.$$

(iii) *For one petite set Z,*

$$\mathbb{P}_x\{\tau_Z < \infty\} = 1, \qquad x \in Z.$$

(iv) *For some pair (s, ν) satisfying the minorization condition (1.7) with $s \in \mathbb{F}^+$,*

$$\nu Hs := \sum_{t=0}^{\infty} \nu(K - s \otimes \nu)^t s = 1.$$

If any of these four equivalent conditions hold, then there exists a σ-finite measure μ which is invariant for the kernel P. It is unique in the sense that any σ-finite invariant measure is a constant multiple of the measure given by

$$\mu_\circ(Y) = \nu H\{Y\} = \sum_{t=0}^{\infty} \nu(K - s \otimes \nu)^t\{Y\}, \qquad Y \in \mathbb{F}. \tag{1.10}$$

Proof. We ask the reader to consult [39, 45] for the equivalence of (i) and (ii). Proofs of the remaining equivalences are provided here to illustrate how the various operators come into play. In particular, we will show that $\mu_\circ$ defines an invariant measure.

The implication (iii) $\Longrightarrow$ (i) is proved in two steps. First, on letting $\mathbb{X}_0 := \{x : \mathbb{P}_x\{\tau_Z < \infty\} = 1\}$, we find that this set is *absorbing*. That is,

$$p(\mathbb{X}_0 \mid x) = 1, \qquad x \in \mathbb{X}_0.$$

It then follows that this set is *full*: $\psi(\mathbb{X}_0^c) = 0$. Hence, from ψ-a.e. initial condition, the set Z is visited infinitely often with probability one. It follows that $G(x, Z) = \infty$ for $x \in \mathbb{X}_0$, and ψ-irreducibility then requires that $G(x, Z) = \infty$ for *all* x.

We now establish a similar identity for arbitrary $Y \in \mathbb{F}^+$. Since Z is petite, for any such Y we can find $\epsilon > 0$ such that

$$K(x, Y) \geq \epsilon \mathbf{1}_Z(x), \qquad x \in \mathbb{X}.$$

This can also be written $K\mathbf{1}_Y \geq \epsilon \mathbf{1}_Z$, and hence, for $x \in \mathbb{X}$,

$$\begin{aligned} \infty &= \epsilon G\mathbf{1}_Z \\ &\leq GK\mathbf{1}_Y = -\mathbf{1}_Y + G\mathbf{1}_Y \leq G\mathbf{1}_Y. \end{aligned}$$

We have thus shown that (iii) implies recurrence.

Conversely, if the chain is recurrent, take any $Z_1 \in \mathbb{F}^+$, and define

$$\mathbb{X}_0 := \{x : \mathbb{P}_x\{\tau_{Z_1} < \infty\} = 1\}; \qquad Z := Z_1 \cap \mathbb{X}_0 \,.$$

The set $\mathbb{X}_0$ is absorbing, so we do find that, for $x \in Z$,

$$\mathbb{P}_x\{\tau_Z < \infty\} = \mathbb{P}_x\{\tau_{Z_1} < \infty\} = 1.$$

This shows that (iii) holds with this set Z, and establishes the implication (i) $\Longrightarrow$ (iii).

We now show that (iv) implies recurrence. To avoid dealing with potentially infinite sums, let $\lambda > 0$, and define the kernels

$$H_\lambda(x) = \sum_0^\infty \lambda^{-n-1}(K - s \otimes \nu)^n \qquad G_\lambda(x) = \sum_0^\infty \lambda^{-n-1} K^n.$$

Note that $H_1 = H$ and $G_1 = G$ are the potential kernels introduced in the introduction.

We denote $\alpha_\lambda = \nu(H_\lambda s)$, and $\beta_\lambda = \nu(G_\lambda s)$. It is obvious that β_λ is finite for each $\lambda > 1$. An application of Lemma 1.1 shows that $\alpha_\lambda \leq 1$ for all $\lambda \geq 1$.

Applying the kernel $\lambda^{-1}K$ to the function $H_\lambda s$ gives,

$$\begin{aligned} \lambda^{-1} K H_\lambda s &= \lambda^{-1}(K - s \otimes \nu) H_\lambda s + (s \otimes \nu) H_\lambda s \\ &= H_\lambda s - \lambda^{-1}(1 - \alpha_\lambda) s. \end{aligned}$$

Iterating this equation we find, for $\lambda > 1$,

$$H_\lambda s - \lambda^{-1}(1 - \alpha_\lambda) \sum_0^{n-1} \lambda^{-i} K^i s = \lambda^{-n} K^n H_\lambda s \to 0, \qquad n \to \infty.$$

This shows that $H_\lambda s = (1 - \alpha_\lambda) G_\lambda s$, and hence that $\alpha_\lambda = (1 - \alpha_\lambda)\beta_\lambda$ for $\lambda > 1$. Letting $\lambda \downarrow 1$ and applying the monotone convergence theorem we see that

$$\beta_1 = \frac{\alpha_1}{1 - \alpha_1}.$$

This shows that $\alpha_1 = 1$ if and only if $\beta_1 = \infty$.

It remains to show that an infinite value for β_1 is equivalent to recurrence. If $\beta_1 = \nu G s = \infty$, it then follows that

$$\int G(x, dy)s(y) = \infty,$$

for ψ-a.e. $x \in \mathbb{X}$ $[\nu]$, and we can extend this to all x by ψ-irreducibility. Let $Y \in \mathbb{F}^+$ be arbitrary, and let $\delta = \nu(Y)$. We can assume by Proposition 1.1 that $\delta > 0$. From the minorization condition we have, $K\mathbf{1}_Y \geq \delta s$, and hence the bound on G gives,

$$\infty = \delta^{-1} \int G(x, dy)s(y) \leq G(x, Y) - \mathbf{1}_Y(x).$$

We deduce that the chain is recurrent as required. The proof that recurrence implies (iv) is identical.

To establish invariance of $\mu_\circ$, first apply the kernel $(K - s \otimes \nu)$ to $\mu_\circ$ on the right to obtain,

$$\mu_\circ(K - s \otimes \nu) = \sum_{t=0}^{\infty} \nu(K - s \otimes \nu)^{t+1} = \mu_\circ - \nu.$$

Now by recurrence and (iv) we have $\mu_\circ(s) = 1$, which shows that $\mu_\circ$ is K-invariant: $\mu_\circ K = \mu_\circ$. Using the identity

$$PK = KP = \beta^{-1}K + (1 - \beta^{-1})I, \tag{1.11}$$

we conclude that $\mu_\circ$ is P-invariant. ∎

The invariant measure given in (1.10) will be finite, so the chain $\boldsymbol{x}$ is positive recurrent, provided that the *mean* return time to a petite set Z is bounded:

$$\sup_{x \in Z} \mathbb{E}_x[\tau_Z] < \infty. \tag{1.12}$$

In terms of the variables used in the previous proof, this is equivalent to requiring that $\alpha_1' < \infty$, where the prime denotes the left derivative of α with respect to λ (see discussion surrounding equation (5.6) of [45]).

While these definitions lead to an elegant theory, in practice one can typically take $\mathbb{X}_0 = \mathbb{X}$ in (ii). In this case the chain is called *Harris*, and it is called *positive Harris* if there is also an invariant probability measure. The chains we consider next exhibit a far stronger form of stability.

1.2.3 $\boldsymbol{c}$-Regularity and Lyapunov functions

The next level of stability that we consider is related to steady state performance, which moves us closer to the average cost optimality criterion. Suppose that $c\colon \mathbb{X} \to [1, \infty)$ is a measurable function on the state space, and suppose that the chain is ψ-irreducible.

Definition 1.4

(i) *A set $S \in \mathbb{F}$ is called* ***c-regular*** *if for any $Y \in \mathbb{F}^+$,*

$$\sup_{x \in S} \mathbb{E}_x \Big[\sum_{t=0}^{\tau_Y - 1} c(x_t) \Big] < \infty.$$

(ii) *The Markov chain is called c-regular if the state space $\mathbb{X}$ admits a countable covering by c-regular sets.*

A c-regular chain is automatically positive Harris, and using (1.10) we see that a c-regular chain possesses an invariant probability measure μ satisfying $\mu(c) := \int c(x)\, \mu(dx) < \infty$. The following result is a consequence of the f-Norm Ergodic Theorem of [39, Theorem 14.0.1].

Theorem 1.2 *Assume that $c \colon \mathbb{X} \to [1, \infty)$ and that $\boldsymbol{x}$ is c-regular. Then, for any measurable function g which satisfies*

$$\sup_{x \in \mathbb{X}} \Big(\frac{|g(x)|}{c(x)} \Big) < \infty,$$

the following ergodic theorems hold for any initial condition:

$$\textit{(i)} \qquad \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} g(x_t) \;=\; \mu(g), \qquad [\mathbb{P}_x] - a.s..$$

$$\textit{(ii)} \quad \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_x[g(x_t)] \;=\; \mu(g).$$

■

One approach to establishing c-regularity is through the following extension of Foster's criterion, or Lyapunov's second method. In general, such approaches involve the construction of a function V on the state space, taking positive values, such that $V(x_t)$ is in some sense decreasing whenever the state x_t is 'large'. In our context this decreasing property can be formulated as follows: Find a function $V \colon \mathbb{X} \to \mathbb{R}_+$ and a constant $\bar{J} \in \mathbb{R}_+$ such that

$$PV\,(x) := \mathbb{E}[V(x_{t+1}) \mid x_t = x] \le V(x) - c(x) + \bar{J}, \qquad x \in \mathbb{X}. \tag{1.13}$$

When this bound holds, we say that V is a *Lyapunov function*.

However, for this to imply any form of stability, the difference $c(x) - \bar{J}$ must be positive for 'large' x. We say that c is

Definition 1.5

near-monotone *if the sublevel set $Z_\eta := \{x \in \mathbb{X} : c(x) \le \eta\}$ is petite for any $\eta < \|c\|_\infty$. The supremum norm $\|c\|_\infty$ may be infinite.*

norm-like *if the sublevel set Z_η is a pre-compact subset of the metric space $\mathbb{X}$ for any η.*

Related assumptions on c are used in [1, 5, 39, 42].

Theorem 1.3 *Assume that* $c\colon \mathbb{X} \to [1,\infty)$ *is near-monotone, and suppose that* $\bar{J} < \|c\|_\infty$. *Then,*

(i) *If there exists a finite, positive-valued solution* V *to the inequality (1.13), then there exists* $d_0 < \infty$ *such that for each* $Y \in \mathbb{F}^+$,

$$\mathbb{E}_x\Big[\sum_{t=0}^{\tau_Y} c(x_t)\Big] \le d_0 V(x) + d(Y), \qquad x \in \mathbb{X}, \tag{1.14}$$

where $d(Y) < \infty$ *is a constant. Hence, each of the sublevel sets* $Z_n = \{x : V(x) \le n\}$ *is c-regular, and the process itself is c-regular.*

(ii) *If the chain is c-regular, then for any c-regular set* $Z \in \mathbb{F}^+$, *the function*

$$V^*(x) = \mathbb{E}_x\Big[\sum_{t=0}^{\tau_Z} c(x_t)\Big], \qquad x \in \mathbb{X}, \tag{1.15}$$

is a near-monotone solution to (1.13).

■

Proof. First observe that the bound (1.13) is equivalent to the drift condition $PV_0 \le V_0 - c + b\mathbf{1}_Z$, where Z is petite: if (1.13) holds, we can take $V_0 = d_0 V$ and $b = d_0\bar{J}$, with d_0 sufficiently large. The result is then an immediate consequence of [39, Theorem 14.2.3]. ■

1.3 PERFORMANCE

For a function $c\colon \mathbb{X} \to [1,\infty)$ we define the average cost by

$$J := \limsup_{n\to\infty} \frac{1}{n}\,\mathbb{E}_x\Big[\sum_{t=0}^{n-1} c(x_t)\Big].$$

We have seen that, for a c-regular chain, the average cost is finite and independent of x, with $J = \mu(c)$. Here we examine the relationship between c and J through Poisson's equation.

1.3.1 Poisson's equation

This functional equation originated in the analysis of partial differential equations: Assuming that f is some given function on $\mathbb{R}^n$, the equation is written

$$\Delta h = -f$$

where h is an unknown function on $\mathbb{R}^n$, and Δ is the Laplacian. The probabilistic interpretation of this equation becomes evident when one realizes that Δ is the generator for a Brownian motion on $\mathbb{R}^n$ - a similar equation can be posed for any Markov process in continuous time. When time is discrete, we define the generator as $\Delta = P - I$, and the Poisson equation then takes on

the exact same form. If we take $f = c - \mu(c)$, then Poisson's equation can be written,

$$Ph\,(x) := \mathbb{E}_x[h(x_{t+1}) \mid x_t = x] = h(x) - c(x) + J, \qquad x \in \mathbb{X}, \tag{1.16}$$

where $J = \mu(c)$.

Motivation for looking at this equation is provided by our prior stability analysis. First note that the drift inequality (1.13) suggests a simple approach to obtaining performance bounds. By iterating this equation one obtains,

$$0 \le \mathbb{E}_x[V(x_n)] \le V(x) - \sum_{t=0}^{n-1} \mathbb{E}_x[c(x_t)] + n\bar{J}. \tag{1.17}$$

Dividing by n and letting $n \to \infty$ then gives the upper bound, $J \le \bar{J}$. The question then is, by choosing V carefully can we get a tight upper bound? The answer is yes, provided the chain is c-regular, and in this case, the minimal upper bound $\bar{J}_*$ is evidently $\mu(c)$, with μ equal to the invariant probability for the chain. We see in Theorem 1.4 that this 'optimal Lyapunov function' is precisely the solution to Poisson's equation.

Equation (1.16) resembles a version of the Lyapunov drift inequality (1.13). However, the function h cannot play the role of a Lyapunov function unless it is positive-valued, or at least bounded from below. This cannot be expected in general, as the following example demonstrates.

Consider the Markov chain on $\mathbb{X} = \mathbb{N}$ with transition probabilities,

$$p(x+1 \mid x) = \begin{cases} \alpha & x > 0; \\ \beta & x < 0; \\ \frac{1}{2} & x = 0. \end{cases} \qquad p(x-1 \mid x) = \begin{cases} \beta & x > 0; \\ \alpha & x < 0; \\ \frac{1}{2} & x = 0. \end{cases}$$

We assume that $\alpha + \beta = 1$, and that $\alpha < \beta$. The latter condition ensures that $\boldsymbol{x}$ is a c-regular T-chain, where c can be taken as any polynomial function of x.

Define $h(x) = x/(\beta - \alpha)$. We can compute the conditional expectation,

$$Ph\,(x) = p(x+1 \mid x)h(x+1) + p(x-1 \mid x)h(x-1) = -\text{sign}\,(x).$$

The function $c(x) = 1 + \text{sign}(x)$, $x \in \mathbb{X}$, is bounded, non-negative, and has steady state mean equal to one. The function h is the associated solution to Poisson's equation,

$$Ph = h - c + 1.$$

It is not bounded from below.

A solution to Poisson's equation *will* be bounded from below under suitable conditions on the chain, and the function c. One such condition is the norm-like assumption, or the milder near-monotonicity condition for c. The following result surveys the relevant consequences of c-regularity, and introduces the form of the Poisson equation that we will analyze in the remainder of this chapter. These results are taken from [42], following [46, 22].

Theorem 1.4 *Assume that* $c\colon \mathbb{X} \to [1,\infty)$ *and that* $\boldsymbol{x}$ *is* c*-regular. Then,*

(i) *There exists a measurable function* $h\colon \mathbb{X} \to \mathbb{R}$ *satisfying (1.16), where* $J = \mu(c)$.

(ii) *One solution to (1.16) may be expressed,*

$$h = \frac{\beta}{1-\beta}[H - I]\bar{c}, \tag{1.18}$$

where $\bar{c} = c - \mu(c)$, $\beta < 1$ *is used in the definition (1.3) of the kernel* K, *and the pair* (s,ν) *is petite with* $s \in \mathbb{F}^+$.

(iii) *Suppose moreover that the function* c *is near-monotone. Then the solution (1.18) is uniformly bounded from below,* $\inf_{x\in\mathbb{X}} h(x) > -\infty$. *It is essentially unique in the following sense: If* h' *is any function on* $\mathbb{X}$ *which is uniformly bounded from below, and solves the Poisson inequality*

$$Ph\,(x) \le h(x) - c(x) + J, \qquad x \in \mathbb{X},$$

with $J = \mu(c)$, *then there exists a constant* k' *such that*

$$\begin{aligned} h'(x) &\ge h(x) + k', \qquad x \in \mathbb{X}; \\ h'(x) &= h(x) + k', \qquad \psi - a.e.\ x \in \mathbb{X}. \end{aligned}$$

(iv) *If* V *is any solution to (1.13) with* $\bar{J} < \|c\|_\infty$ *and* c *near-monotone, then the solution (1.18) satisfies the uniform upper bound, for some* $d_0 < \infty$,

$$h(x) \le d_0(V(x) + 1), \qquad x \in \mathbb{X}.$$

Proof. To prove (i) and (ii), consider first the function

$$h_0(x) = H\bar{c} := \sum_{t=0}^{\infty} (K - s\otimes\nu)^t \bar{c}\,(x), \qquad x \in \mathbb{X}.$$

The construction of $\mu_\circ$ in (1.10) gives

$$\nu(h_0) = \mu_\circ(\bar{c}) = \mu_\circ(\mathbb{X})\mu(\bar{c}) = 0.$$

This immediately gives,

$$Kh_0\,(x) = (K - s\otimes\nu)h_0\,(x) = h_0(x) - \bar{c}(x).$$

That is, h_0 solves the Poisson equation for the kernel K. By applying the identity (1.11) we see that

$$h := \frac{\beta}{1-\beta}(h_0 - \bar{c}) = \frac{\beta}{1-\beta} Kh_0,$$

solves the Poisson equation for original transition kernel P.

To prove (iii), define h_0' by

$$h_0' := \frac{1-\beta}{\beta} h' + \bar{c}.$$

The resolvent equation (1.11) combined with the bound in (iii) gives,

$$\frac{1-\beta}{\beta} Kh' \le \frac{1-\beta}{\beta} h' - K\bar{c},$$

from which it follows that $Kh_0' \le h_0' - \bar{c}$. This implies that the quantity $\nu(h_0')$ is finite:

$$s(x)\nu(h_0') \le Kh_0' \le h_0'(x) - \bar{c}(x), \qquad x \in \mathbb{X}.$$

By adding a constant to h' we can and will assume that $\nu(h_0') = 0$. We then have,

$$(K - s \otimes \nu) h_0' \le h_0' - \bar{c}$$

We have assumed that h' (and hence h_0') is bounded from below. By iterating the last inequality, it follows that for some $L < \infty$,

$$\begin{aligned} -L &\le (K - s \otimes \nu)^n h_0' \\ &\le h_0' - \sum_0^{n-1} (K - s \otimes \nu)^i \bar{c} \end{aligned}$$

We conclude that

$$h_0' \ge h_0 - L.$$

Letting $u = h_0' - h_0$, we see that u is bounded from below, and it is *superharmonic*: $Pu \le u$. These properties imply the desired result (see [39, p. 414] or [42]).

The proof of (iv) is similar to (iii). We first establish a bound of the form,

$$PV \le V - \epsilon c + bs,$$

with $\epsilon > 0$, $b < \infty$. We can move to the resolvent to obtain an analogous bound,

$$(K - s \otimes \nu) V \le KV \le V - \epsilon_1 c + b_1 s,$$

and on iterating we find that

$$\epsilon_1 Hc \le b_1 Hs + V \le b_1 + V.$$

In Lemma 1.1 we have shown that Hs is everywhere bounded by unity. The bound in (iv) immediately follows. ∎

1.3.2 Simulation

We have now seen that the Poisson equation has a direct role in performance evaluation since an approximation of the solution h will lead to an approximation of $J = \mu(c)$ using (1.17). With some structure imposed on the model this idea does lead to algorithms for computing bounds on J. For example, this is the essence of the main results in [35, 36], where performance bounds are obtained in the network scheduling problem. If the cost is linear, and if any of the linear programs constructed in these references admits a feasible solution, then the solution to Poisson's equation is approximated by a pure quadratic function.

Perhaps the most obvious approach to estimating J is through Monte-Carlo simulation via

$$\widehat{J}_n = \frac{1}{n} \sum_{0}^{n-1} c(x_t), \qquad n \in \mathbb{N}.$$

The Poisson equation again plays an important role in analysis, and in the generation of more efficient simulation approaches.

The effectiveness of the Monte Carlo method depends primarily on the magnitude of the Central Limit Theorem variance, also known as the time-average variance. Under suitably strong recurrence conditions on the Markov chain this can be expressed

$$\gamma_c^2 = \lim_{n \to \infty} \mathbb{E}_x \Big[\Big(\frac{1}{\sqrt{n}} \sum_{0}^{n-1} \bar{c}(x_t) \Big)^2 \Big] .$$

An alternative expression for the time-average variance is computed through the formula

$$\gamma_c^2 = \mu(h^2) - \mu((Ph)^2) = 2\mu(h\bar{c}) - \mu(\bar{c}^2), \tag{1.19}$$

with h any solution to Poisson's equation [39, eq. 17.50].

There are many variants of the simple Monte-Carlo estimate, some of which may have far smaller variance. After all, if $\{\Delta_t : t \geq 0\}$ is any sequence of random variables satisfying $\frac{1}{n} \sum_0^{n-1} \Delta_t \to 0$, $n \to \infty$, then the modified estimator,

$$\widehat{J}_n^{\Delta} = \frac{1}{n} \sum_{0}^{n-1} (c(x_t) + \Delta_t), \qquad n \in \mathbb{N},$$

is another consistent estimator of J. An *optimal* choice for Δ_t is computed using the solution h to Poisson's equation (1.16): by setting

$$\Delta_t^* = Ph\,(x_t) - h(x_t),$$

we obtain a time-average variance of *zero*. Of course, computing Δ_t^* involves a computation of J, so this approach is nonsensical! If however an approximation g to h can be found, then the choice $\Delta_t = Pg\,(x_t) - g(x_t)$ will lead to reduced variance if the approximation is sufficiently tight [24, 25].

We will discover such approximations when we attempt to solve some optimization problems below, and hence we will have many candidates for the function g.

1.3.3 Examples

In this chapter we develop two general examples: the linear state space model, and a family of network models. In this section we look at some special cases without control. Controlled linear systems, and controlled network models are considered as examples in the final section of this chapter.

The linear state space model is defined through the multi-dimensional recursion,

$$x_{t+1} = Ax_t + Fw_{t+1}, \qquad t \in \mathbb{N}, \tag{1.20}$$

where $x_t \in \mathbb{R}^d, w_t \in \mathbb{R}^q$, A is a $d \times d$ matrix, and F is a $d \times q$ matrix. We assume that $\boldsymbol{w}$ is i.i.d., and that $\boldsymbol{w}$ is a Gaussian process with mean zero, and covariance I. That is, $w_t \sim N(0, I)$, where I is the identity matrix.

The *controllability matrix* is the $d \times (dq)$ matrix

$$\mathcal{C} := [A^{d-1}F | A^{d-2}F | \cdots | AF | F],$$

where the bar denotes concatenation of matrices. The pair (A, F) is called *controllable* if the matrix $\mathcal{C}$ has rank d [37]. The process is ψ-irreducible with ψ equal to Lebesgue measure if the pair (A, F) is controllable. To see why, note that the state at time d can be written,

$$x_d = A^d x_0 + \mathcal{C} w_1^d$$

where $w_1^d = (w_1^{\mathrm{T}}, \ldots, w_d^{\mathrm{T}})^{\mathrm{T}}$. It follows that x_d itself is Gaussian with mean $A^d x_0$, and covariance given by

$$\Sigma_d = \mathcal{C}\mathcal{C}^{\mathrm{T}}.$$

The covariance is full rank if the model is controllable, and it follows that $P^t(x, \cdot)$ is equivalent to Lebesgue measure for any x, and any $t \geq d$. By continuity of the model it is easy to check that (1.7) holds with s continuous, and ν equal to normalized Lebesgue measure on an open ball in $\mathbb{R}^d$. We conclude that $\boldsymbol{x}$ is a T-chain if the controllability condition holds. A sample path from a particular two dimensional linear model is shown in Figure 1.1. When the state is large, the sample path behavior appears almost deterministic.

To find a stochastic Lyapunov function V with $c(x) = \frac{1}{2}x^T Q x$, first solve the *Lyapunov equation*,

$$A^T M A = M - Q. \tag{1.21}$$

If $M > 0$ (M is positive definite) then $V(x) = \frac{1}{2}x^T M x$ is a solution to (1.13). Direct calculations show that the function V is also the essentially unique solution to Poisson's equation, with $J = \frac{1}{2}\mathrm{trace}\,(F^T M F)$.

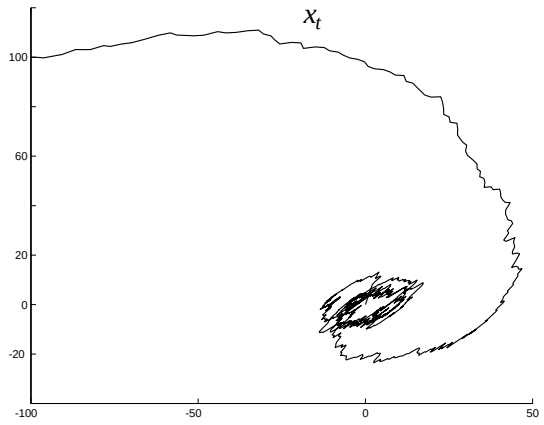

Figure 1.1 A sample path of the linear state space model with a 'large' initial condition $x_0 = \binom{-100}{100}$.

For the nonlinear state space model

$$x_{t+1} = F(x_t, w_{t+1}), \qquad t \in \mathbb{N},$$

the ψ-irreducibility condition can still be verified under a nonlinear controllability condition called *forward accessibility* [39, Chapter 7]. The construction of a Lyapunov function is however far more problem-specific.

Over the past five years there has been much research on algorithmic methods for constructing Lyapunov functions for *network models*. One is based upon linear programming methods, and is similar to the Lyapunov equation (1.21) used for linear state space models [35]. We describe here a recent approach based upon a fluid model [41, 42]. As an example we consider here the simplest case: An uncontrolled M/M/1 queue.

When the arrival stream is renewal, and the service times are i.i.d., then the waiting time for a simple queue can be modeled as a Markov chain with state space $\mathbb{X} = \mathbb{R}_+$. The dynamics take the form of a one dimensional linear state space model, where the state space is constrained to the positive half line. The queue length process is itself a Markov process in the special case where the service times and interarrival times are exponentially distributed. By applying *uniformization* (i.e. sampling the process appropriately - see [38]), the queue length process $\boldsymbol{x}$ obeys the recursion

$$x_{t+1} = x_t + (1 - I_{t+1})\pi(x_t) + I_{t+1}, \qquad t \in \mathbb{N},$$

where $\boldsymbol{I}$ is a Bernoulli, i.i.d. random process: $\lambda = \mathbb{P}(I_t = 1)$ is the arrival rate, and $\mu = \mathbb{P}(I_t = -1)$ is the service rate. The function π plays the role of a 'policy', where in this simple example we take $\pi(x) = \mathbf{1}(x > 0)$. Time has been normalized so that $\lambda + \mu = 1$.

To construct a Lyapunov function, first note that stability is a 'large state' property, so it may pay to consider the process starting from a large initial

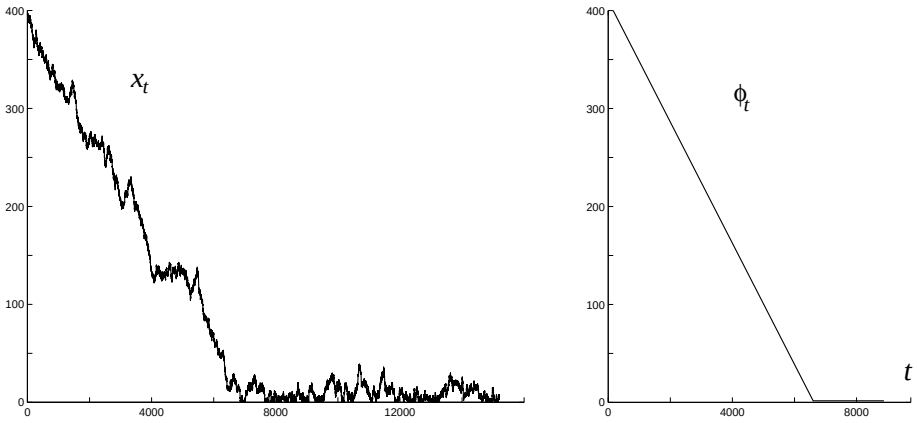

Figure 1.2 On the left is a sample path x_t of the M/M/1 queue with $\rho = \lambda/\mu = 0.9$, and $x_0 = 400$. On the right is a solution to the differential equation $\dot{\varphi} = (-\mu + \lambda)\pi(\varphi)$ starting from the same initial condition.

condition. In the left hand side of Figure 1.2 we see one such simulation. As was seen in the linear model, when the initial condition is large the behavior of the model is roughly deterministic.

Suppose we take the cost function $c(x) = 1 + x$. To construct a Lyapunov function we would ideally like to compute the expected sum given in (1.15), with S equal to some finite set, perhaps $S = \{0\}$. While this is computable for the M/M/1 queue, such computation can be formidable for more complex network models. However, consider the right hand side of Figure 1.2 which shows a sample path of the deterministic fluid, or leaky bucket model. This satisfies the differential equation $\dot{\varphi} = (-\mu + \lambda)\pi(\varphi)$, where π is again equal the indicator function of the strictly positive real axis. The behavior of the two processes look similar when viewed on this large spatial/temporal scale. It appears that a good approximation is

$$\begin{aligned} V(x) &:= \int_0^\infty \varphi(t)\, dt, \quad \varphi(0) = x, \\ &= \tfrac{1}{2}\frac{x^2}{\mu - \lambda}. \end{aligned} \tag{1.22}$$

If we apply the transition kernel P to V we find, for $x \geq 1$,

$$\begin{aligned} PV\,(x) &= \lambda V(x+1) + \mu V(x-1) \\ &= \frac{1}{2(\mu - \lambda)}\big(\lambda(x+1)^2 + \mu(x-1)^2\big) \\ &= V(x) - x + \frac{1}{2(\mu - \lambda)}, \end{aligned}$$

while for $x = 0$ we have,

$$PV\,(x) = \frac{\lambda}{2(\mu - \lambda)} \leq V(x) - x + \frac{1}{2(\mu - \lambda)}$$

That is, we see that this approach works: The stochastic Lyapunov criterion (1.13) does hold with this function V derived from the fluid model, where $\bar{J} = (2(\mu - \lambda))^{-1}$, under the stability condition that $\rho = \lambda/\mu < 1$. The actual steady state mean of $c(x) = x$ is given by $J = \lambda(\mu - \lambda)^{-1}$, which is indeed upper-bounded by $\bar{J}$.

What about the more exact Poisson's equation? Can the fluid model be used to approximate a solution?

With the cost function $c(x) = 1 + x$, the Poisson equation for the M/M/1 queue becomes

$$Ph = \lambda h(x+1) + \mu h((x-1)^+) = h(x) - c(x) + J.$$

One solution is given by

$$h(x) = \frac{x^2 + x}{2(\mu - \lambda)},$$

which is similar in form to the fluid value function given in (1.22).

For a general class of network models it can be shown that the value function for the fluid model and the solution to Poisson's equation are roughly equal for large x in the sense that

$$h(x) = V(x)(1 + o(1)),$$

where the term $o(1) \to 0$ as $x \to \infty$. Some results of this type are described in Section 1.6.2.

The M/M/1 queue illustrates the difficulties one faces in using simulation: Using (1.19) we can show that the time-average variance constant γ_c^2 is of order $(1-\rho)^{-4}$ in this example since pth moments for the M/M/1 queue are of order $(1-\rho)^{-p}$ [25].

With this background we are now ready to turn to MDP models.

1.4 THE AVERAGE COST OPTIMALITY EQUATION

We now assume that there is a control sequence taking values in the action space $\mathbb{A}$ which influences the behavior of $\boldsymbol{x}$. The state space $\mathbb{X}$ and the action space $\mathbb{A}$ are assumed to be locally compact, separable metric spaces, and we continue to let $\mathbb{F}$ denote the Borel σ-field on $\mathbb{X}$. Associated with each $x \in \mathbb{X}$ is a non-empty and closed subset $\mathbb{A}(x) \subseteq \mathbb{A}$ whose elements are admissible actions when the state process x_t takes the value x. The set of admissible state-action pairs $\{(x,a) : x \in \mathbb{X}, a \in \mathbb{A}(x)\}$ is assumed to be a measurable subset of the product space $\mathbb{X} \times \mathbb{A}$.

The transitions of $\boldsymbol{x}$ are governed by the conditional probability distributions $\{p(Y|x,a) : Y \in \mathbb{F}, x \in \mathbb{X}, a \in \mathbb{A}(x)\}$ which describe the probability that the next state is in Y, given that the current state is x, and the current action chosen is a. These are assumed to be probability measures on $\mathbb{F}$ for each state-action pair (x,a), and measurable functions of (x,a) for each $Y \in \mathbb{F}$.

We recall the following definitions:

Definition 1.6

(i) *A **nonrandomized policy** ϕ is a sequence of measurable functions ϕ_n, $n \in \mathbb{N}$, from H_n to $\mathbb{A}$ such that $\phi_n(x_0, a_0, \ldots x_{n-1}, a_{n-1}, x_n) \in \mathbb{A}(x_n)$.*

(ii) *If for each n this function ϕ_n depends only on x_n, then the policy ϕ is called **Markov**. The set of all Markov policies is denoted Π^M.*

(iii) *If ϕ is Markov, and if there is a fixed function π such that $\phi_n = \pi$ for all n, then the policy is called **stationary**. We denote by Π^S the set of all stationary policies.*

*For convenience, we extend the notation by writing $\pi \in \Pi^S$ when π is a measurable function from $\mathbb{X}$ to $\mathbb{A}$ which defines a stationary policy ϕ. The function π is called a **feedback law**.*

For any $\pi \in \Pi^S$, the state process $\boldsymbol{x}^\pi := \{x_t^\pi : t \geq 0\}$ is a Markov chain on $(\mathbb{X}, \mathbb{F})$ with stationary transition probabilities.

We do not consider randomized policies. This is without loss of generality since we can always redefine the MDP model so that the action space $\mathbb{A}$ is replaced with the *space of probability measures on* $\mathbb{A}$. An ordinary policy for the new MDP model is equivalent to a randomized one for the original model.

We shall write

$$P_\pi^t f = \int_{\mathbb{X}} f(y) p^t(dy|x, \pi(x)), \qquad f \in L_\infty, t \geq 1,$$

for the semigroup of kernels corresponding to a policy $\pi \in \Pi^S$, and we let K_π denote the corresponding resolvent kernel. We continue to use the operator-theoretic notation,

$$P_\pi^t h\,(x) := \mathbb{E}^\pi[h(x_t^\pi) \mid x_0 = x].$$

In the remainder of this section we describe consequences of the average cost optimality equation, and develop criteria for existence of solutions. These results are based on [42, 44]. More background may be found in [29, 5, 28, 1, 49].

1.4.1 Regular and optimal policies

We suppose that a one-step *cost function* $c\colon \mathbb{X} \times \mathbb{A} \to [1, \infty)$ is given. Other chapters in this volume consider a reward function r. Throughout this chapter we will take $c = -r$, so that the optimization problem becomes one of *cost minimization*. We assume below that c satisfies a near-monotone condition so that the results of Section 1.2 and 1.3 may be applied. We will use freely terminology that was introduced in these sections. In particular, we refer the reader to Definition 1.5 for a definition of near-monotonicity, and Definition 1.4 for c-regularity and related topics.

The steady state average cost is denoted

$$J(\phi, x) := \limsup_{n \to \infty} \frac{1}{n}\, \mathbb{E}_x \Big[\sum_{t=0}^{n-1} c(x_t^\phi, \phi_t) \Big].$$

For any $\phi \in \Pi^S$, defined by a state feedback law π, we define $c_\pi\colon \mathbb{X} \to \mathbb{R}$ as the function given by $c_\pi(x) = c(x, \pi(x))$, $x \in \mathbb{X}$. A policy ϕ will be called

Definition 1.7

regular *if* $\phi \in \Pi^S$, *and* $\boldsymbol{x}^\phi$ *is a* c_π*-regular Markov chain.*

s-optimal *if* $\phi \in \Pi^S$ *and*

$$J(\phi, x) \leq J(\phi', x), \qquad \phi' \in \Pi^S, x \in \mathbb{X}.$$

m-optimal *if* $\phi \in \Pi^M$ *and*

$$J(\phi, x) \leq J(\phi', x), \qquad \phi' \in \Pi^M, x \in \mathbb{X}.$$

Many of the results below concern conditions which guarantee the existence of a regular, s-optimal policy π_*. In this case $J_* = J(\pi_*, x)$ is independent of x.

When J_* is independent of x, and the optimization criterion is *cost minimization*, the associated *average cost optimality equation* (ACOE) is given as follows. The function h_* is known as the *relative value function*..

$$J_* + h_*(x) = \min_{a \in \mathbb{A}(x)} [c(x,a) + P_a h_*(x)] \tag{1.23}$$

$$\pi_*(x) = \arg\min_{a \in \mathbb{A}(x)} [c(x,a) + P_a h_*(x)], \qquad x \in \mathbb{X}. \tag{1.24}$$

If a stationary policy π_*, a measurable function h_*, and a constant J_* exist which solve (1.23,1.24), then typically the policy π_* is optimal (see for example [1, 5, 27, 49, 55] for a proof of this and related results).

Theorem 1.5 *Suppose that* (J_*, h_*, π_*) *solve (1.23,1.24). Assume moreover that, for any* $x \in \mathbb{X}$, *and any* $\pi \in \Pi^S$ *satisfying* $J(\pi, x) < \infty$,

$$\frac{1}{n} P_\pi^n h_*(x) \to 0, \qquad n \to \infty. \tag{1.25}$$

Then π_* *is an s-optimal control, and* J_* *is the optimal cost, in the sense that*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_x[c_{\pi_*}(x_t^{\pi_*})] = J_*,$$

and $J(\pi, x) \geq J_*$ *for all stationary policies* π, *and all initial states* x. ∎

The assumption (1.25) is unfortunate, but examples show that some additional conditions on h_* are required (see e.g. page 87 of [5], Chapter 7 of [55], or the examples in [1, 15, 49, 51]). The following result gives a condition implying (1.25) which is often verifiable in practice, as we shall see in Section 1.6.1 and Section 1.6.2.

Suppose that $\phi \in \Pi^S$ is defined by a feedback law π, suppose that the controlled chain $\boldsymbol{x}^\pi$ is ψ_π-irreducible, and that $\mu_\pi(c_\pi) = \int c_\pi(x)\mu_\pi(dx)$ is finite. Let $Z_\pi \in \mathbb{F}$ denote any fixed c_π-regular set for which $\mu_\pi(Z_\pi) > 0$. We then define

$$V_\pi(x) = \mathbb{E}_x\Big[\sum_{t=0}^{\tau_\pi - 1} c_\pi(x_t^\pi)\Big], \qquad x \in \mathbb{X}, \tag{1.26}$$

where $\tau_\pi = \tau_{Z_\pi}$ is the first entrance time to Z_π. Since $\mu_\pi(Z_\pi) > 0$, the function V_π is μ_π-a.e. finite-valued [39, Theorem 14.2.5]. Note that, by [39, Theorem 14.2.3], the particular c_π-regular set Z_π chosen is not important. If Z_π^1 and Z_π^2 give rise to functions V_π^1 and V_π^2 of the form (1.26), then for some constant $\gamma \geq 1$,

$$\gamma^{-1} V_\pi^1(x) \leq V_\pi^2(x) \leq \gamma V_\pi^1(x), \qquad x \in \mathbb{X}.$$

The following result is taken from [42]. We show how the assumptions are verified in Section 1.6.

Theorem 1.6 *Suppose that the optimality equation (1.23,1.24) holds for* (J_*, h_*, π_*)*, with* h_* *bounded from below. For any other* $\pi \in \Pi^S$ *assume that*

(a) *The function*

$$K_\pi c_\pi = (1-\beta) \sum_{t=0}^{\infty} \beta^t P_\pi^t c_\pi$$

is norm-like, and the Markov chain $\boldsymbol{x}^\pi$ *is a* T*-chain.*

(b) *There exists some constant* $d_0 = d_0(\pi) < \infty$ *such that,*

$$|h_*(x)| \leq d_0 V_\pi(x), \qquad x \in \mathbb{X}. \tag{1.27}$$

Then π_* *is a regular, s-optimal policy.* ∎

Theorem 1.6 asserts that a solution to the ACOE will yield an optimal policy provided h_* is 'small enough'. This motivates the construction of a solution through the formulation of a *minimal* relative value function. We consider this approach next.

1.4.2 Existence of solutions

The ACOE is a generalization of Poisson's equation. To formulate sufficient conditions for a solution, we generalize the operators H and K which formed the basis of analysis in Section 1.3. In an attempt to simplify the development, we assume here that the cost $c(\,\cdot\,,\,\cdot\,)$ is a function of x only. Analogous results hold for general cost functions.

Randomization is required to define the resolvent kernel for a Markov policy. Let $\{\xi_i : i \geq 1\}$ denote an i.i.d. sequence of geometrically distributed random variables which is independent of the controlled chain. We assume that for some $0 < \beta < 1$,

$$\mathbb{P}(\xi_i = k) = (1-\beta)\beta^k, \qquad k \geq 0,\ i \geq 1.$$

For any Markov policy, we then write,

$$K_\phi(x, Y) = \mathbb{P}^\phi\{x_{\xi_1} \in Y \mid x_0 = x\}, \qquad x \in \mathbb{X},\ Y \in \mathbb{F}.$$

This coincides with the previous definition (1.3) when the policy ϕ is stationary.

A potential kernel is then defined in analogy with (1.5) by sampling the chain at the renewal events $t_k = \sum_{i=1}^{k} \xi_i$. Let $\boldsymbol{y}$ denote the sampled, controlled chain $y_k = x_{t_k}$, $k \geq 0$. A Markov policy for $\boldsymbol{y}$ is expressed as a *sequence* of Markov policies for $\boldsymbol{x}$, via

$$\boldsymbol{\phi} := (\phi_1, \phi_2, \dots), \qquad \phi_i \in \Pi^M, i \geq 1.$$

We let $\boldsymbol{\Pi}^M$ denote the set of all such sequences, and we let $\boldsymbol{\Pi}^S \subset \boldsymbol{\Pi}^M$ denote the set of trivial sequences constructed from a *stationary* policy. That is, $\boldsymbol{\phi} \in \boldsymbol{\Pi}^S$ if $\boldsymbol{\phi} = (\phi, \phi, \dots)$, with $\phi \in \Pi^S$.

Given a policy $\boldsymbol{\phi} \in \boldsymbol{\Pi}^M$, the transition probabilities are given by

$$\mathbb{P}(y_{k+\ell} \in Y \mid y_k = x) = [K_{\phi_{k+1}} \cdots K_{\phi_{k+\ell}}](x, Y), \qquad x \in \mathbb{X},\ Y \in \mathbb{F},\ k \geq 0,\ \ell \geq 1.$$

For a fixed $s \colon \mathbb{X} \to]0, 1[$ and a probability measure ν on $\mathbb{F}$ we define

$$\begin{aligned} M_n^{\boldsymbol{\phi}} &:= (K_{\phi_1} - s \otimes \nu) \cdots (K_{\phi_n} - s \otimes \nu), \qquad n \geq 1. \\ H^{\boldsymbol{\phi}} &:= I + \sum_{n=1}^{\infty} M_n^{\boldsymbol{\phi}}. \end{aligned}$$

This again agrees with our earlier definition of H when $\boldsymbol{\phi} \in \boldsymbol{\Pi}^S$. To ensure positivity of these kernels we impose a minorization condition below.

These operators immediately give formulae for a candidate solution to the ACOE. We first define a candidate average cost:

$$\eta(\boldsymbol{\phi}) := \inf\Big(\eta : \nu H^{\boldsymbol{\phi}}(c - \eta) \leq 0\Big), \qquad \boldsymbol{\phi} \in \boldsymbol{\Pi}^M; \tag{1.28}$$

$$\eta_* := \inf\Big(\eta(\boldsymbol{\phi}) : \boldsymbol{\phi} \in \boldsymbol{\Pi}^M\Big). \tag{1.29}$$

The construction of a solution to Poisson's equation in Theorem 1.4 leads to the following definitions: For any $x \in \mathbb{X}$,

$$h_*^0(x) := \inf\Big(H^{\boldsymbol{\phi}}(c - \eta_*)(x) : \boldsymbol{\phi} \in \boldsymbol{\Pi}^M\Big). \tag{1.30}$$

$$h_*(x) := \Big(\frac{\beta}{1-\beta}\Big) \inf\Big(K_\phi h_*^0(x) : \phi \in \Pi^M\Big). \tag{1.31}$$

A candidate s-optimal policy is then,

$$\pi_*(x) := \arg\min_{a \in \mathbb{A}} \int_{\mathbb{X}} p(dy \mid x, a) h_*(y), \qquad x \in \mathbb{X}. \tag{1.32}$$

The following assumptions ensure that these functions are well defined, and that the function h_* is bounded from below. Provided these assumptions hold, and that there exists at least one 'stabilizing policy', we find that the triple (J_*, h_*, π_*) solves the ACOE with $J_* = \eta_*$.

(A1) The infimums in (1.30,1.31) exist, and admit measurable solutions h_*^0 and h_*. Moreover, the minimum in (1.32) exists point-wise to form a stationary (measurable) policy π_*.

(A2) There exists a norm-like function $\underline{c}\colon \mathbb{X} \to \mathbb{R}_+$ such that,

$$K_\phi c \geq \underline{c}, \qquad \phi \in \Pi^M.$$

(A3) There exists a continuous function $s\colon \mathbb{X} \to]0,1[$, and a probability measure ν on $\mathbb{F}$, such that

$$K_\phi(x, Y) \geq s(x)\nu(Y), \qquad x \in \mathbb{X},\ Y \in \mathbb{F},\ \phi \in \Pi^M.$$

The measurability assumption in (A1) is, surprisingly, the most subtle of these three conditions. A strong Feller assumption, that $p(h \mid x, a)$ is a continuous function of (x, a) for a sufficiently large class of functions h will imply that h_* is continuous, and hence measurable. The existence of a measurable solution π_* in (1.32) will require further conditions (such as compactness of $\mathbb{A}(x)$ for all x).

Lemma 1.2 *If (A1)-(A3) hold then, whenever the invariant probability μ_π exists,*

$$\eta_* \leq \mu_\pi(c), \qquad \pi \in \Pi^S.$$

Proof. This follows from the construction of $\mu(\,\cdot\,) = \mu_\circ(\,\cdot\,)/\mu_\circ(\mathbb{X})$ given in (1.10), the definition (1.29), and the minorization condition (A3). ∎

Thus, it is not surprising that π_* is an s-optimal policy:

Theorem 1.7 *Suppose that Assumptions (A1)-(A3) are satisfied, and suppose that there exists one regular policy $\pi_0 \in \Pi^S$ with average cost $\mu_{\pi_0}(c) < \infty$.*
Then the following hold:

(a) *The triple (J_*, h_*, π_*) solve the ACOE, where $J_* = \eta_*$, and the policy π_* is regular and s-optimal.*

(b) *The function h_* is uniformly bounded from below:*

$$\inf_{x \in \mathbb{X}} h_*(x) > -\infty.$$

(c) *If (h'_*, J_*) is any other solution to (1.23), with h_* uniformly bounded from below, then there exists a constant k' such that*

$$h'_*(x) \begin{cases} = h_*(x) + k' & \text{for almost every } x \\ \geq h_*(x) + k' & \text{for every } x. \end{cases}$$

Proof. We will just prove (a). This and the remaining parts are similar to the proof of Theorem 1.4. A complete proof in the countable state space setting is given in [44].

We first show that h_*^0 solves the ACOE for the resolvent. From the definition of η_* we find that

$$\nu(h_*^0) \leq 0. \tag{1.33}$$

For any $\phi \in \Pi^M$, $\boldsymbol{\phi} \in \boldsymbol{\Pi}^M$, we can apply (1.33) and the definition of h_*^0 to give,

$$\begin{aligned} K_\phi h_*^0 &\le (K_\phi - s \otimes \nu) h_*^0 \\ &\le (K_\phi - s \otimes \nu) H^{\boldsymbol{\phi}}(c - \eta_*) \\ &= H^{\boldsymbol{\phi}^{[1]}}(c - \eta_*) - c + \eta_* \end{aligned}$$

where $\boldsymbol{\phi}^{[1]} := (\phi, \phi_1, \phi_2, \dots)$.

Infimizing over all $\phi \in \Pi^M$, $\boldsymbol{\phi} \in \boldsymbol{\Pi}^M$ gives the upper bound,

$$\inf_\phi \Big(K_\phi h_*^0(x) \Big) \le h_*^0(x) - c(x) + \eta_*, \qquad x \in \mathbb{X}.$$

We also know that h_*^0 is finite-valued by the regularity assumption imposed on π_0: With $\phi_0 \in \Pi^S$ equal to the stationary policy defined by π_0, the following bound follows from minimality of h_*^0,

$$h_*^0 \le H^{\boldsymbol{\phi}^0}(c - \eta_*), \qquad \boldsymbol{\phi}^0 := (\phi_0, \phi_0, \dots) \in \boldsymbol{\Pi}^S.$$

We now turn to the function h_*. Exactly as in the uncontrolled case (see the proof of Theorem 1.4), we can translate from the resolvent to the original chain to obtain,

$$\begin{aligned} P_{\pi_*} h_*(x) &:= \inf_{a \in \mathbb{A}(x)} P_a \Big(\inf_{\phi \in \Pi^M} K_\phi h_*^0(x) \Big) \\ &\le h_*(x) - c(x) + \eta_*, \qquad x \in \mathbb{X}. \end{aligned}$$

It follows that $\mu_{\pi_*}(c) \le \eta_*$, and then by minimality of η_* we must have equality. Hence by Theorem 1.4, the above is an equality for μ_{π_*}-a.e. x. By minimality of h_* and Theorem 1.4 (iv), it must be an equality for all x. ∎

The literature on average cost optimal control is filled with counter-examples. It is of some interest then to see why Theorem 1.7 does not fall into any of these traps. Consider first counter-examples 1 and 2 of [51, p. 142]. In each of these examples the MDP is completely non-irreducible in the sense that

$$\mathbb{P}(x_t^\pi < x_0^\pi) = 0, \qquad t \ge 1, \pi \in \Pi^S.$$

It is clear then from the cost structure that the bound (A3) on the resolvent cannot hold in this case.

Another example is given in the Appendix of [51] in which a version of (A3) is directly assumed! However, the cost is not unbounded, and is in fact designed to favor large states.

The assumptions (A2) and (A3) together imply that the center of the state space, as measured by the cost criterion, possesses some minimal amount of irreducibility. If either the unboundedness condition or the accessibility condition is relaxed, so that the process is non-irreducible on a set where the cost is low, then we see from these counter-examples that optimal stationary policies may not exist.

In the remainder of this chapter we preserve these three assumptions. They will be generalized sightly when we consider algorithms in the next section.

1.5 ALGORITHMS

Value iteration and policy iteration are two well-known algorithms for constructing optimal policies. The value iteration algorithm, or VIA, is a version of successive approximation. The policy iteration algorithm, or PIA, first proposed in [31], may be interpreted as a version of the Newton-Raphson method. We find that the PIA is more easily analyzed under the assumptions we impose even though the algorithm is considerably more complex than value iteration. The ease of analysis is a result of the hard work already taken care of in Section 1.3.

Although complex, the PIA may converge extremely quickly when properly normalized. See [20, 42] for application in the communication and network areas.

The results below are taken from [42, 10]. Related work on algorithms may be found in [30, 7, 54, 8].

1.5.1 Value iteration

The ACOE (1.23) can be viewed as a fixed point equation in the variables (h_*, J_*). By ignoring the constant term, and applying successive approximation to this fixed point equation, we obtain the VIA. Suppose that the positive-valued function V_n is given. Then the stationary policy π_n is defined as

$$\pi_n(x) = \arg\min_{a \in \mathbb{A}(x)} [P_a V_n(x) + c(x,a)], \qquad x \in \mathbb{X},$$

and one then defines

$$V_{n+1}(x) = c_{\pi_n}(x) + P_{\pi_n} V_n(x) = \min_{a \in \mathbb{A}(x)} (P_a V_n(x) + c_a(x)), \tag{1.34}$$

which then makes it possible to compute the next policy π_{n+1} by re-starting the algorithm.

This is in fact the standard dynamic programming approach to constructing a finite horizon optimal policy since for each n we may write,

$$V_n(x) = \min\Big(\mathbb{E}_x^{\phi} \Big[\sum_{t=0}^{n-1} c(x_t, a_t) + V_0(\Phi(n)) \Big] : \phi \in \Pi^M \Big). \tag{1.35}$$

We see in (1.35) that the initial condition V_0 plays the role of a terminal penalty function.

The initialization V_0 should be chosen with care. For a countable state space model, a poor choice (such as $V_0 \equiv 0$) can lead to policies for which the controlled chain is transient [10]. We assume in Theorem 1.8 below that at least one regular policy π_{-1} exists, and that the function V_0 serves as a Lyapunov function: for some constant $\bar{J} < \infty$,

$$P_{\pi_{-1}} V_0 \le V_0 - c_{\pi_{-1}} + \bar{J}. \tag{1.36}$$

The existence of a pair (V_0, π_{-1}) satisfying (1.36) is a natural stabilizability assumption on the model, and we find below that this initialization ensures that the VIA generates stabilizing policies.

To simplify notation we define $c_n = c_{\pi_n}$, $P_n = P_{\pi_n}$, and we define the resolvent for the nth policy by

$$K_n := (1-\beta)\sum_{t=0}^{\infty} \beta^t P_n^t, \qquad n \geq 0, \tag{1.37}$$

where $\beta \in]0,1[$ as before. We let $\mathbb{E}^n$ denote the expectation operator induced by the stationary policy π_n.

Let ν denote some fixed probability measure on $\mathbb{F}$. We define, for each n, the normalized value function, and the incremental cost,

$$h_n(x) = V_n(x) - \nu(V_n); \qquad \gamma_n(x) = V_{n+1}(x) - V_n(x), \qquad x \in \mathbb{X}, n \in \mathbb{N}. \tag{1.38}$$

From the definitions, for each n we have the familiar looking identity $P_n h_n = h_n - c_n + \gamma_n$.

Defining $\bar{J}_n = \sup_x \gamma_n(x) \leq \infty$, we obtain the following solution to (1.13):

$$P_n V_n \leq V_n - c_n + \bar{J}_n. \tag{1.39}$$

Under (A1)-(A3), and with an initial condition satisfying (1.36), we find that the $\{\bar{J}_n\}$ are finite valued, and non-increasing. The assumptions below are almost identical to (A1)-(A3) in Section 1.4.

(VIA1) For each n, if the VIA yields a value function $V_n : \mathbb{X} \to \mathbb{R}_+$, then for each $x \in \mathbb{X}$ the minimization

$$\pi_n(x) := \arg\min_{a \in \mathbb{A}(x)} \big[c(x,a) + P_a V_n(x)\big]$$

exists, and admits a measurable solution π_n.

(VIA2) There exists a norm-like function $\underline{c} : \mathbb{X} \to \mathbb{R}_+$ such that

$$K_n c_n(x) \geq \underline{c}(x), \qquad x \in \mathbb{X}, n \in \mathbb{N}.$$

(VIA3) There is a fixed probability ν on $\mathbb{F}$, a $\delta > 0$, and an initial value function V_0 with the following property: For each $n \geq 1$, if the VIA yields the value function V_n, then for any policy π_n given in (VIA1),

$$K_n(x, Y) \geq \delta \nu(Y) \qquad x \in S, Y \in \mathbb{F}, \tag{1.40}$$

where S denotes the pre-compact set

$$S = \{x : \underline{c}(x) \leq 2\bar{J}\}. \tag{1.41}$$

The following result is largely taken from [10].

Theorem 1.8 *Suppose that (VIA1)–(VIA3) hold. Assume moreover that the initialization V_0 satisfies (1.36). Then,*

(i) *Each of the policies $\{\pi_i : i \in \mathbb{N}\}$ is regular.*

(ii) *The upper bounds $\{\bar{J}_n\}$ are decreasing:*

$$\bar{J}_0 \geq \bar{J}_1 \geq \cdots \geq \bar{J}_n \geq \cdots;$$

(iii) *The sequence $\{h_n\}$ is uniformly bounded from below.*

Proof. The minimization in the value iteration algorithm immediately leads to the bound $P_n\gamma_n \geq \gamma_{n+1}$. From this we deduce by induction that the $\bar{J}_n$ are finite and decreasing: The initialization of the induction relies on the assumption that the initial condition V_0 satisfies (1.36). This then proves (ii).

To establish (i), note first that the following bound on the resolvent follows from (1.39):

$$K_n V_n \leq V_n - \frac{\beta}{1-\beta} K_n c_n + \frac{\beta}{1-\beta}\bar{J}_n. \tag{1.42}$$

This inequality is a version of (1.13) since $V_n \geq 0$, and we have established that $\bar{J}_n$ is finite. Applying Theorem 1.3 and using (VIA2,VIA3) for the kernel K_n, we see that the Markov chain with transition kernel K_n is c-regular. This implies (i).

To prove (iii) note first of all that $h_n(x) \geq -\nu(V_n) > -\infty$ for all x. It remains to obtain a bound independent of n. For any n we have

$$K_n h_n \leq h_n - K_n c_n + \bar{J} \leq h_n + \bar{J}\mathbf{1}_S$$

Letting $s = \delta \mathbf{1}_S$ we then obtain,

$$(K_n - s \otimes \nu) h_n \leq h_n + \bar{J}\delta^{-1} s,$$

and by iteration, for any N,

$$-\nu(V_n)(K_n - s\otimes\nu)^N \mathbf{1} \leq (K_n - s\otimes\nu)^N h_n \leq h_n + \bar{J}\delta^{-1} \sum_{i=0}^{N-1} (K_n - s\otimes\nu)^i s.$$

By c_n-regularity of the nth chain it follows that $(K_n - s\otimes\nu)^N \mathbf{1}\,(x) \to 0$ as $N \to \infty$ for any x. This and Lemma 1.1 then gives the bound

$$0 \leq h_n + \bar{J}\delta^{-1} \sum_{i=0}^{\infty} (K_n - s\otimes\nu)^i s \leq h_n + \bar{J}\delta^{-1}.$$

∎

Convergence of the algorithm is subtle. This is not surprising since it is rare in optimization to prove global convergence of successive approximation. The countable state space case is considered in [10] where it is shown that (VIA1), (VIA2), and a strengthening of (VIA3) do imply convergence of $\{h_n\}$ to a solution of the ACOE. To generalize this result to general state spaces it may be necessary to impose a blanket stability condition as in [29], or the stronger stability assumption imposed in [14, 56].

1.5.2 Policy iteration

The PIA, which is again a recursive algorithm for generating stationary policies, follows naturally as a refinement of the VIA. We saw that the value iteration algorithm generates regular policies because we have established in Section 1.5.1 the drift inequality,

$$P_{n-1} V_{n-1} \leq V_{n-1} - c_{n-1} + \bar{J}_{n-1}.$$

From this bound we discovered easily that the next policy π_n has cost bounded by $J(\pi_n, x) \le \bar{J}_{n-1}$, $x \in \mathbb{X}$. We have seen that there are an infinite number of solutions to drift inequalities of this form, and some give better bounds than others. The *optimal* solution is the solution to Poisson's equation, since this gives the minimal possible value for $\bar{J}$. On replacing the function V_{n-1} by the solution to Poisson's equation in the VIA recursion (1.34) one obtains precisely the PIA.

To give a precise description of the algorithm, suppose that at the $(n-1)$th stage of the algorithm a stationary policy π_{n-1} is given, and assume that h_{n-1} satisfies the Poisson equation

$$P_{n-1}h_{n-1} = h_{n-1} - c_{n-1} + J_{n-1},$$

where $P_{n-1} = P_{\pi_{n-1}}$, $c_{n-1}(x) = c_{\pi_{n-1}}(x) = c(x, \pi_{n-1}(x))$, and J_{n-1} is a constant (equal to the steady state cost with this policy).

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

$$\pi_n(x) = \arg\min_{a \in \mathbb{A}(x)} \big[c(x,a) + P_a h_{n-1}(x)\big]. \tag{1.43}$$

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

Our analysis of the PIA is based on the pair of equations

$$P_n h_n = h_n - \bar{c}_n; \tag{1.44}$$

$$P_n h_{n-1} = h_{n-1} - \bar{c}_n + \gamma_n, \tag{1.45}$$

where $\bar{c}_n = c_n - J_n$, and γ_n is now *defined* through (1.45). From the minimization (1.43) 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

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

Combining these two equations gives the upper bound $\gamma_n(x) \le J_{n-1} - J_n$, $x \in \mathbb{X}$, which shows that the PIA automatically generates solutions to (1.13).

As was the case with the value iteration algorithm, much of the analysis of [42] focuses on $\{K_n\}$ rather than $\{P_n\}$, as given in (1.37). To invoke the algorithm we must again ensure that the required minimum exists.

(PIA1) For each n, if the PIA yields a triplet $(J_{n-1}, h_{n-1}, \pi_{n-1})$ which solve Poisson's equation

$$P_{n-1}h_{n-1} = h_{n-1} - c_{n-1} + J_{n-1},$$

with h_{n-1} bounded from below, then for each $x \in \mathbb{X}$ the minimization

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

exists, and admits a measurable solution π_n.

(PIA2) There exists a norm-like function $\underline{c}\colon \mathbb{X} \to \mathbb{R}_+$ such that for the policies π_n obtained through the PIA,

$$K_n c_n(x) \geq \underline{c}(x), \qquad x \in \mathbb{X}, n \in \mathbb{N}.$$

(PIA3) There is a fixed probability ν on $\mathbb{F}$, a $\delta > 0$, and an initial regular policy π_0 with the following property: For each $n \geq 1$, if the PIA yields a triplet $(J_{n-1}, h_{n-1}, \pi_{n-1})$ with h_{n-1} bounded from below, then for any policy π_n given in (PIA1),

$$K_n(x, Y) \geq \delta\nu(Y) \qquad x \in S, Y \in \mathbb{F}, \tag{1.46}$$

where S denotes the pre-compact set

$$S = \{x : \underline{c}(x) \leq 2J_0\}. \tag{1.47}$$

Under Assumptions (PIA1)-(PIA3), the algorithm produces stabilizing policies recursively. A proof of Theorem 1.9 may be found in [42].

Theorem 1.9 *Suppose that (PIA1)-(PIA3) hold, and that the initial policy π_0 is regular. Then for each n the PIA admits a solution (J_n, h_n, π_n) such that π_n is regular, and the sequence of relative value functions $\{h_n\}$ defined in (1.18) satisfy,*

(i) *For some constant $N < \infty$,*

$$\inf_{x \in \mathbb{X}, n \geq 0} h_n(x) > -N;$$

(ii) *There exists $h_\infty\colon \mathbb{X} \to \mathbb{R}$ such that*

$$\lim_{n \to \infty} h_n(x) = h_\infty(x), \qquad x \in \mathbb{X};$$

(iii) *There exists $b_1, b_2 < \infty$ such that*

$$-N \leq h_\infty(x) \leq b_1 h_0(x) + b_2, \qquad x \in \mathbb{X}.$$

■

Now that we know that $\{h_n\}$ is point-wise convergent to a function h_∞, we can show that the PIA yields a solution to the ACOE. We let π_∞ denote a solution to

$$\pi_\infty(x) = \arg\min_{a \in \mathbb{A}(x)} P_a h(x), \qquad x \in \mathbb{X}. \tag{1.48}$$

Theorem 1.10 is similar to Theorem 4.3 of [27] which requires a related continuity condition. Weaker conditions are surely possible for a specific application.

Theorem 1.10 *Suppose that (PIA1)-(PIA3) hold, and that the initial policy π_0 is regular. Assume in addition that*

(i) *The function π_∞ in (1.48) can be chosen to form a stationary policy.*

(ii) *The function $c\colon \mathbb{X}\times\mathbb{A} \to [1,\infty)$ is continuous, and the functions $(P_a h_n\,(x) : n \geq 0)$ and $P_a h\,(x)$ are continuous in (a,x).*

(iii) *For each $x \in \mathbb{X}$, the function $c(x,\,\cdot\,)$ is norm-like on $\mathbb{A}$.*

(iv) *The initial condition h_0 satisfies,*

$$\lim_{n\to\infty} \frac{1}{n} P_\pi^n h_0\,(x) = 0,$$

for any $\pi \in \Pi^S$, and any $x \in \mathbb{X}$ for which $J(\pi,x) < \infty$.

Then,

(a) *The PIA produces a sequence of solutions (J_n, h_n, π_n) such that $\{J_n, h_n\}$ is point-wise convergent to (J_∞, h_∞). The triple $(J_\infty, h_\infty, \pi_\infty)$ is a solution to the ACOE.*

(b) *The policy π_∞ is c_{π_∞}-regular, and s-optimal. Consequently, for any initial condition $x \in \mathbb{X}$,*

$$J(\pi_\infty, x) = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_x^{\pi_\infty}[c_\pi(x_t^{\pi_\infty})] = \mu_{\pi_\infty}(c_\pi) = J_\infty.$$

■

We now illustrate the theory with some general examples.

1.6 EXAMPLES

1.6.1 Linear models

The controlled linear state space model is defined through the recursion,

$$x_{t+1} = Ax_t + Ba_t + Fw_{t+1}, \qquad t \in \mathbb{N}, \tag{1.49}$$

where $w_t \in \mathbb{R}^q$, $x_t \in \mathbb{R}^d$, and $a_t \in \mathbb{R}^p$. As before we assume that $w \sim N(0,I)$, and that (A,F) is controllable (see the discussion following (1.20)).

Since $\boldsymbol{w}$ is i.i.d., then this is a Markov decision process with transition function

$$p(Y \mid x,a) = \mathbb{P}(w_1 + Ax + Ba \in Y), \qquad x \in \mathbb{X}, Y \in \mathbb{F}, a \in \mathbb{A}.$$

The cost c is taken as the general quadratic,

$$c(x,a) = \tfrac{1}{2} x^T Q x + \tfrac{1}{2} a^T R a, \tag{1.50}$$

with $Q \geq 0$, and $R > 0$. The assumption $c \geq 1$ fails in this example. However, c is positive, so that we can add 1 to the cost function to satisfy the desired lower bound on c and the MDP is essentially unchanged.

The optimization of $J(\pi, x)$ is known as the LQG (linear-quadratic-Gaussian) problem. Under certain conditions on the model, it is known that one may obtain a solution (J_*, h_*, π_*) to the ACOE with h_* quadratic, and π_* linear in x, by solving a Riccati equation [37]. What conditions are required? Why should a solution give rise to an optimal policy?

Assumption (A1) in Section 1.4 requires the existence of measurable solutions to the static optimization problem arising in (1.23,1.24). Since the candidate relative value function is quadratic, this will hold under our assumption that $R > 0$.

The norm-like condition (A2) requires additional assumptions on Q. Let $\sqrt{Q}$ denote any $d \times d$ matrix for which $Q = \sqrt{Q}^{\mathrm{T}}\sqrt{Q}$, and suppose that $(A, \sqrt{Q})$ is *observable*. Algebraically, this means that $(A^T, \sqrt{Q}^T)$ is a *controllable* pair. Physically, it means that the cost will be large whenever the state is large. In [42] it is shown that observability implies (A2), and it is also shown that for any regular policy, the solution to Poisson's equation is bounded from below by a quadratic function of x.

Assumption (A3) will not hold for *any* stationary policy, but if one restricts to policies with bounded growth, say

$$\|\pi(x)\| \leq b_1(\|x\|^2 + 1) \qquad x \in \mathbb{X},$$

then this assumption will hold if F is a $d \times d$ matrix with rank d. This is stronger than the controllability assumption.

Under these conditions it follows from Theorem 1.6 that the linear/quadratic solution (π_*, h_*) to the ACOE does yield an optimal control over the class of all nonlinear feedback laws (i.e., all stationary policies).

Theorem 1.9 recovers known properties of the Newton-Raphson technique applied to the LQG problem, which is precisely the PIA. Suppose the initial policy π_0 in the PIA is linear. One can verify that the solution to Poisson's equation is quadratic. Each subsequent policy is of the form $\pi_n(x) = -K_n x$, for some $p \times n$ matrix K_n, and each subsequent solution to Poisson's equation is quadratic,

$$h_n(x) = h_n(0) + \tfrac{1}{2}x^{\mathrm{T}}\Lambda_n x, \qquad x \in \mathbb{X}.$$

Under the observability condition, it follows from Theorem 1.9 that the matrices $\{\Lambda_n\}$, which are solutions to a *Riccati recursion*, are uniformly bounded in n.

The proof of Theorem 1.9 depends upon a bound of the form,

$$h_n(x) \leq [1 + 2(J_{n-1} - J_n)]h_{n-1}(x) + b(J_{n-1} - J_n), \tag{1.51}$$

where b is a constant (see [42]). Letting $x \to \infty$, it follows that

$$\Lambda_n \leq [1 + 2(J_{n-1} - J_n)]\Lambda_{n-1}, \qquad n \geq 1. \tag{1.52}$$

It is known that the matrices $\{\Lambda_n\}$ are *decreasing*, in the sense that $\Lambda_n - \Lambda_{n-1}$ is positive semidefinite for each $n \geq 1$ [18, 62]. Hence the bound (1.51) is

not tight in the linear model. However, the semi-decreasing property (1.52) is sufficient to deduce convergence of the matrices $\{\Lambda_n\}$ to a finite limiting matrix.

There is no space here to consider the VIA in further detail. We note however that it is well known that the successive approximation procedure generates stabilizing policies for the linear state space model provided the initial policy is stabilizing and linear [18]. Theorem 1.8 shows that it is enough to assume only stability.

1.6.2 Network models

We now apply the general results of Section 1.4 to the scheduling problem for multiclass queuing networks. For simplicity we discuss here only a relatively simple class of network models which can be formulated through an extension of the M/M/1 model. A treatment of general network models is given in [41, 43].

Consider a network composed of d single server stations, indexed by $\sigma = 1, \dots, d$. The network is populated by ℓ classes of customers: Class k customers require service at station $s(k)$. An exogenous stream of customers of class 1 arrive to machine $s(1)$, and subsequent routing of customers is deterministic. 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,

$$x_{t+1} = x_t + \sum_{k=0}^{\ell} I_{t+1}(k)[e^{k+1} - e^k]a_t(k), \qquad t \ge 0, \tag{1.53}$$

where the state process $\boldsymbol{x}$ evolves on the countable state space $\mathbb{X} = \mathbb{N}^\ell$, and $x_t(k)$ denotes the number of class k customers in the system at time t. An example of a two station network is illustrated in Figure 1.3.

The random variables $\{I_t : t \ge 1\}$ are i.i.d. on $\{0,1\}^{\ell+1}$, with

$$\mathbb{P}\{\textstyle\sum_i I_t(k) = 1\} = 1, \text{ and } \mathbb{E}[I_t(k)] = \mu_k.$$

For $1 \le k \le \ell$, 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 $1 \le k \le \ell$ we let e^k denote the kth basis vector in $\mathbb{R}^\ell$, and we set $e^0 = e^{\ell+1} := 0$.

The sequence $\{a_t : t \ge 0\}$ is the control, which takes values in $\mathbb{A} := \{0,1\}^{\ell+1}$. We define $a_t(0) \equiv 1$. The set of admissible control actions $\mathbb{A}(x)$ is defined in an obvious manner: for $a \in \mathbb{A}(x)$,

(i) For any $1 \le k \le \ell$, $a(k) = 0$ or 1;

(ii) For any $1 \le k \le \ell$, $x_k = 0 \Rightarrow a(k) = 0$;

(iii) For any station σ, $0 \le \sum_{k:s(k)=\sigma} a(k) \le 1$;

(iv) For any station σ, $\sum_{k:s(k)=\sigma} a(k) = 1$ whenever $\sum_{k:s(k)=\sigma} x(k) > 0$.

If $a(k) = 1$, then buffer k is chosen for service. Condition (ii) then imposes the physical constraint that a customer cannot be serviced at a buffer if that

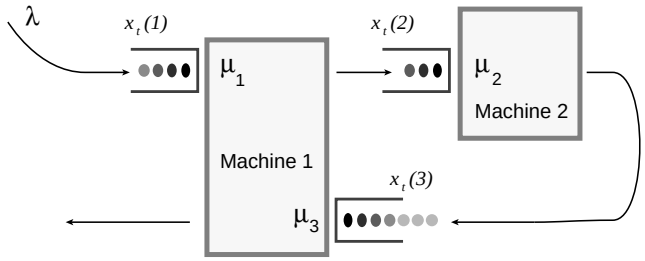

Figure 1.3 A multiclass network with $d = 2$ and $\ell = 3$.

buffer is empty. Condition (iii) means that only one customer may be served at a given instant at a single machine σ.

Since the control is bounded, a reasonable cost function is $c(x, a) = c^T x$, where $c \in \mathbb{R}^\ell$ is a vector with strictly positive entries. For concreteness, we take $c(x, a) = |x| := \sum_k x(k)$. The non-idling condition (iv) is satisfied by any optimal stationary policy with this cost criterion: An inductive proof is given in [41] based upon value iteration.

The controlled transition function has the simple form,

$$\begin{aligned} p(x + e^{k+1} - e^k \mid x, a) &= \mu_k a(k), \qquad 0 \le k \le \ell. \\ p(x \mid x, a) &= 1 - \sum_0^\ell \mu_k a(k) \end{aligned}$$

The accessibility condition (1.7) holds with s everywhere positive, and $\nu = \delta_\theta$, with $\boldsymbol{\theta}$ equal to the empty state $\boldsymbol{\theta} = (0, \dots, 0)^T \in \mathbb{X}$. This follows from the non-idling assumption (iv).

Associated with this network is a *fluid model*. For each initial condition $x_0 \neq 0$, we construct a continuous time process $\varphi^{x_0}(t)$ as follows. If $m = |x_0|$, and if tm is an integer, we set

$$\varphi^{x_0}(t) = \frac{1}{m} x_{mt}.$$

For all other $t \ge 0$, we define $\varphi^{x_0}(t)$ by linear interpolation, so that it is continuous and piecewise linear in t. Note that $|\varphi^{x_0}(0)| = 1$, and that φ^{x_0} is Lipschitz continuous. The collection of all "fluid limits" is defined by

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

where the overbar denotes weak closure in $C(\mathbb{R})$, the space of continuous functions, with the topology of uniform convergence on compact sets. This set of stochastic process of course depends on the particular policy π which has been applied.

Any $\varphi \in \mathcal{L}$ evolves on the state space $\mathbb{R}_+^\ell$ and, for a wide class of scheduling policies, satisfies a differential equation of the form

$$\frac{d}{dt}\varphi(t) = \sum_{k=0}^{\ell} \mu_k [e^{k+1} - e^k] u_t(k), \tag{1.54}$$

where the function u_t is analogous to the discrete control a_t, and satisfies similar constraints (see the M/M/1 queue model described earlier, or [12, 11] for more general examples).

Stability of (1.53) in terms of c-regularity is closely connected with the stability of the fluid model [12, 35, 13]. The fluid model $\mathcal{L}$ is called *L_p-stable* if

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

It is shown in [35] that L_2-stability of the fluid model is equivalent to a form of c-regularity for the network:

Theorem 1.11 *The following stability criteria are equivalent for the network under any non-idling, stationary policy.*

(i) The drift condition (1.13) holds for some function V. The function V is equivalent to a quadratic in the sense that, for some $\gamma > 0$,

$$1 + \gamma|x|^2 \le V(x) \le 1 + \gamma^{-1}|x|^2, \qquad x \in \mathbb{X}. \tag{1.55}$$

(ii) For some quadratic function V,

$$\mathbb{E}_x\Big[\sum_{n=0}^{\sigma_\theta} |x_n|\Big] \;\le\; V(x), \qquad x \in \mathbb{X},$$

where σ_θ is the first entrance time to $\boldsymbol{\theta} = 0$.

(iii) For some quadratic function V and some $\bar{J} < \infty$,

$$\sum_{n=1}^{N} \mathbb{E}_x[|x_n|] \;\le\; V(x) + N\bar{J}, \quad \text{for all } x \text{ and } N \ge 1.$$

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

∎

Using this result it is shown in [42] that when applying policy iteration to a network model, on performing the fluid scaling one obtains a sequence of fluid models which are the solutions of a policy iteration scheme for the fluid model. Moreover, the algorithm convergence to yield a policy which is s-optimal for both the network and its fluid model. The minimal relative value function h_* is equivalent to a quadratic in the sense that, for some constant b_1,

$$b_1^{-1}\|x\|^2 - b_1 \le h_*(x) \le b_1\|x\|^2 + b_1$$

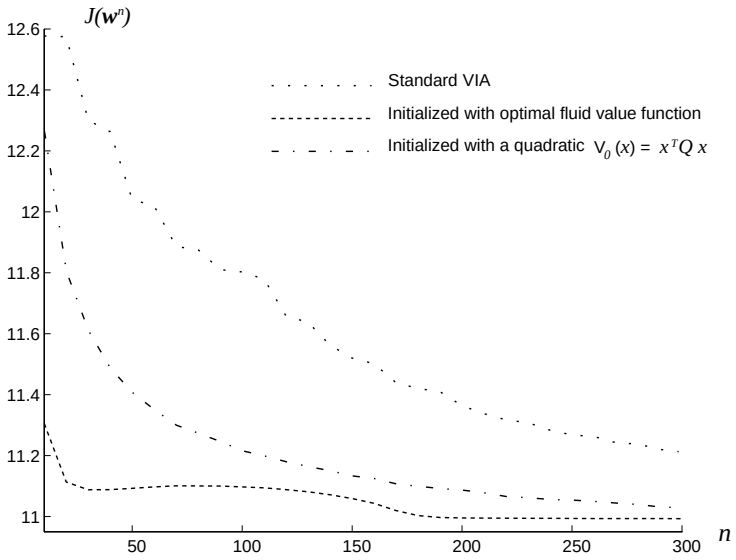

Figure 1.4 Convergence of the VIA with V_0 taken as the value function for the associated fluid control problem, or a pure quadratic function obtained through a linear program.

For any stationary policy, the solution to Poisson's equation is bounded from below by a quadratic.

Let us turn to the VIA: In view of Theorem 1.11, how should we initialize the algorithm? Two possibilities are suggested:

(i) Given the previous analysis of the M/M/1 queue it appears natural to set V_0 equal to the value function for a fluid model,

$$V_*(x) := \min \int_0^\infty |\varphi(t)|\, dt \qquad \varphi(0) = x,\ x \in \mathbb{X},$$

where the minimum is with respect to all policies for the fluid model. One can show that for large x, V_* does approximate the relative value function [41, 42, 25].

(ii) The conclusion that the relative value function is 'nearly quadratic' suggests that we search for a pure quadratic form satisfying (1.13),

$$V_0(x) = x^T Q x, \qquad x \in \mathbb{X}.$$

In [35] a linear program is constructed to compute a quadratic solution to (1.13) for network models, based on prior results of [36, 47].

We conclude with a numerical experiment to show how a careful initialization can dramatically speed convergence of the VIA. We consider the three buffer model illustrated in Figure 1.3 with the following parameters: $\lambda/\mu_2 = 9/10$; $\lambda/\mu_1 + \lambda/\mu_3 = 9/11$; and $\mu_1 = \mu_3$. The optimal value function V_* can be

computed explicitly in this case, and a pure quadratic Lyapunov function can also be computed easily.

Two experiments were performed to compare the performance of the VIA initialized with these two value functions. To apply value iteration the buffer levels were truncated so that $x_i < 45$ for all i. This gives rise to a finite state space MDP with $45^3 = 91,125$ states. The results from two experiments are shown in Figure 1.4. For comparison, data from the standard VIA with $V_0 \equiv 0$ is also given. We have taken 300 steps of value iteration, saving data for $n = 10, \ldots, 300$. The convergence is exceptionally fast in both experiments. Note that the convergence of J_n is *not* monotone in the experiment shown using the fluid value function initialization. However, this initialization leads to fast convergence to the optimal cost $J_* \approx 10.9$.

1.7 EXTENSIONS AND OPEN PROBLEMS

It is hoped that the development in this chapter has suggested to the reader some interesting topics for further research. We list here some areas which have been of interest to the author.

Existence and structure of optimal policies. The results of Section 1.4 are fairly complete, but the setting is special. It appears that there is still much to be done to better understand the structure of optimal policies, and criteria for existence of optimal policies in this general setting.

Continuous time. In this chapter the analysis has been restricted to a resolvent kernel, and the same approach can be followed in continuous time where the resolvent becomes

$$K = \int_0^\infty \beta e^{-\beta t} P^t \, dt$$

with $\beta > 0$. Again one can show that any variable of interest (the invariant measures, solutions to Poisson's equation, or solutions to (1.13)) can be mapped between the resolvent and the continuous time process. Further discussion may be found in [42, 55].

Geometric ergodicity and risk sensitive control. The risk sensitive control criterion is given via

$$J_\gamma(\pi, x) := \limsup_{n\to\infty} \frac{1}{n} \log\Big(\mathbb{E}_x\Big[\exp\Big(\gamma \sum_{t=0}^{n-1} c_\pi(x_t^\pi)\Big)\Big]\Big).$$

where the 'risk factor' γ is assumed to be a small, positive number in the *risk-averse* case.

Models of this sort were first considered in [3, p. 329] and in [32, 52]. This control problem has attracted more recent attention because of the interesting connections between risk sensitive control and game theory [33, 21, 62].

Under a norm-like condition on the model it can be shown that when this cost is finite valued, the Markov chain exhibits a strong form of stability known

as geometric ergodicity [2, 6]. Conversely, such stability assumptions imply that the cost is finite, and ensure that an optimal policy does exist [23, 9].

Our present understanding of the optimization problem for Markov chains on an infinite state space is currently weak, and this appears to be an area worthy of further study.

Simulation. The use of simulation will become increasingly important in both evaluating and synthesizing policies. Much of the burden of finding an optimal policy surrounds the solution of Poisson's equation, for which now there are several simulation based algorithms such as temporal difference learning. There are also simulation based versions of both value and policy iteration (see [4, 57, 34]).

We have remarked that high variance can make simulation impractical. The use of the fluid value function is one promising approach to variance reduction for network models [25], and related techniques may prove useful in the development of simulation-based optimization algorithms.

Complexity. This has always been one of the most challenging issues in optimal control. Markovian models are frequently too 'fine-grained' to be useful in optimization. One solution then is to seek some form of aggregation. For general MDP models one can directly discretize the state space to obtain a finite state space model.

This is an area in which the most relevant research will most likely focus on a specific application. In the case of network models, either fluid models or Brownian motion models provide approaches to aggregation which deserve further study.

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. Control Optim.*, 31:282–344, 1993.

[2] S. Balaji and S.P. Meyn. Multiplicative ergodicity and large deviations for an irreducible Markov chain. *Stochastic Process. Appl.*, 90:123–144, 2000.

[3] R. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, NJ, 1957.

[4] Bertsekas, D., Tsitsiklis, J. Neuro-Dynamic Programming, Athena Scientific, Belmont, MA, 1996.

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

[6] V. Borkar and S.P. Meyn. Risk Sensitive Optimal Control: Existence and Synthesis for Models with Unbounded Cost. To appear, *Mathematics of O.R.*, 2000.

[7] R. Cavazos-Cadena. Value iteration in a class of communicating Markov decision chains with the average cost criterion. *SIAM J. Control and Optimization*, 34:1848–1873, 1996.

[8] 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. Applied Probability*, 33:986–1002, 1996.

[9] R. Cavazos-Cadena and E. Fernandez-Gaucherand. Controlled Markov chains with risk-sensitive criteria: Average cost, optimality equations, and optimal solutions. Mathematical Methods of Operations Research, 49:299-324, 1999.

[10] R-R. Chen and S.P. Meyn. Value iteration and optimization of multiclass queueing networks. *Queueing Systems*, 32:65–97, 1999.

[11] J. Dai and G. Weiss. Stability and instability of fluid models for certain re-entrant lines. *Mathematics of Operations Research*, 21(1):115–134, February 1996.

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

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

[14] R. Dekker. *Denumerable Markov Decision Chains: Optimal Policies for Small Interest Rates*, PhD thesis, University of Leiden, Leiden, the Netherlands, 1985.

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

[16] C. Derman. Denumerable state MDPs. *Ann. Amth. Statist.*, 37:1545–1554, 1966.

[17] D. Down, S. P. Meyn, and R. L. Tweedie. Geometric and uniform ergodicity of Markov processes. *Ann. Probab.*, 23(4):1671–1691, 1996.

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

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

[20] E. A. Feinberg, Ya. A. Kogan and A. N. Smirnov. Optimal Control by the Retransmission Probability in Slotted ALOHA Systems. *Performance Evaluation*, 5:85-96, 1985.

[21] W.H. Fleming and W.M. McEneaney. Risk-sensitive control and differential games. volume 84 of *Lecture Notes in Control and Info. Sciences*, pages 185–197. Springer-Verlag, Berlin; New York, 1992.

[22] P.W. Glynn and S.P. Meyn. A Liapunov bound for solutions of Poisson equation. *Annals of Prob.*, 24:916–931, 1996.

[23] D. Hernández-Hernández and S.I. Marcus. Risk sensitive control of Markov processes in countable state space. *Systems Control Lett.*, 29:147–155, July 1996. correction in *Systems and Control Letters*, 34:105-106, 1998.

[24] Henderson, S. G. *Variance Reduction Via an Approximating Markov Process.* Ph.D. thesis. Department of Operations Research, Stanford University. Stanford, California, USA, 1997.

[25] S.G. Henderson and S.P. Meyn. Variance reduction for simulation in multiclass queueing networks. *submitted to the IIE Transactions on Operations Engineering: special issue honoring Alan Pritsker on simulation in industrial engineering*, 1999.

[26] J. Humphrey D. Eng and S.P. Meyn. Fluid network models: Linear programs for control and performance bounds. In J. Cruz J. Gertler and M. Peshkin, editors, *Proceedings of the 13th IFAC World Congress*, volume B, pages 19–24, San Francisco, California, 1996.

[27] O. Hernández-Lerma and J. B. Lasserre. Discrete time Markov control processes I. Springer-Verlag, New York, 1996.

[28] 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.*, 28:29–46, 1991.

[29] A. Hordijk. *Dynamic Programming and Markov Potential Theory. Math. Centre Tracts, Mathematical Centrum, Amsterdam*, 2nd ed., 1977.

[30] A. Hordijk and M. L. Puterman. On the convergence of policy iteration. *Math. Op. Res.*, 12:163–176, 1987.

[31] R. A. Howard. *Dynamic Programming and Markov Processes.* John Wiley and Sons/MIT Press, New York, NY, 1960.

[32] R.A. Howard and J.E. Matheson. Risk-sensitive Markov decision processes. *Management Sci.*, 8:356–369, 1972.

[33] D. H. Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Trans. Automat. Control*, AC-18:124–131, 1973.

[34] V. R. Konda and V. S. Borkar. Actor-critic-type learning algorithms for Markov decision processes. *SIAM J. Control and Optimization*, 38:4-123, 1999.

[35] P.R. Kumar and S.P. Meyn. Duality and linear programs for stability and performance analysis queueing networks and scheduling policies. *IEEE Transactions on Automatic Control*, 41(1):4–17, 1996.

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

[37] H. Kwakernaak and R. Sivan. *Linear Optimal Control Systems.* Wiley-Interscience, New York, NY, 1972.

[38] S. Lippman. Applying a new device in the optimization of exponential queueing systems. *Operations Research*, 23:687–710, 1975.

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

[40] S. P. Meyn and R. L. Tweedie. Stability of Markovian processes III: Foster-Lyapunov criteria for continuous time processes. *Adv. Appl. Probab.*, 25:518–548, 1993.

[41] S.P. Meyn. Stability and optimization of multiclass queueing networks and their fluid models. In *Mathematics of Stochastic Manufacturing Systems*, Lectures in Applied Mathematics, Vol. 33. Proc. AMS-SIAM Summer Seminar in Applied Mathematics June 17-22, 1996, Williamsburg, Virginia. G. George Yin and Qing Zhang (Eds.). American Mathematical Society, Providence, 1997,

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

[43] S.P. Meyn. Feedback regulation for sequencing and routing in multiclass queueing networks. *SIAM J. Control and Optimization*, to appear, 1999.

[44] S.P. Meyn. Algorithms for optimization and stabilization of controlled Markov chains. *Sadhana*, 24:1-29, 1999.

[45] E. Nummelin. *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge University Press, Cambridge, 1984.

[46] E. Nummelin. On the Poisson equation in the potential theory of a single kernel. *Math. Scand.*, 68:59–82, 1991.

[47] D. Bertsimas, I. C. Paschalidis, and J. N. Tsitsiklis, Optimization of Multiclass Queueing Networks: Polyhedral and Nonlinear Characterizations of Achievable Performance, *Annals of Applied Probability*, 4:43-75, 1994.

[48] J. Perkins. *Control of Push and Pull Manufacturing Systems.* PhD thesis, University of Illinois, Urbana, IL, September 1993. Technical report no. UILU-ENG-93-2237 (DC-155).

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

[50] R. K. Ritt and L. I. Sennott. Optimal stationary policies in general state space Markov decision chains with finite action set. *Mathematics of Operations Research*, 17(4):901–909, November 1993.

[51] S. M. Ross. Applied probability models with optimization applications. Dover books on advanced Mathematics, 1992. Republication of the work first published by Holden-Day, 1970.

[52] U.G. Rothblum. Multiplicative Markov decision chains. *Math. Operations Res.*, 9:6–24, 1984.

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

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

[55] L.I. Sennott. Stochastic Dynamic Programming and the Control of Queueing Systems. *Wiley*, 1999.

[56] F.M. Spieksma. *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*, PhD thesis, University of Leiden, Leiden, the Netherlands, 1990.

[57] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Trans. on Automatic Control*, 42:674-690, 1997.

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

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

[60] G. Weiss. Optimal Draining of a Fluid Re-Entrant Line. In *Stochastic Networks*. Volume 71 of IMA volumes in Mathematics and its Applications, pp. 91-103. Frank Kelly and Ruth Williams, eds. Springer-Verlag, New York, 1995.

[61] G. Weiss. Optimal Draining of Fluid Re-Entrant Lines: Some Solved Examples. In *Stochastic Networks: Theory and Applications*. Volume 4 of Royal Statistical Society Lecture Notes Series , pp. 19-34, F.P. Kelly, S. Zachary and I. Ziedins eds. Oxford University Press, Oxford, 1996.

[62] P. Whittle. *Risk-Sensitive Optimal Control*. John Wiley and Sons, Chichester, NY, 1990.