MARKOV DECISION PROCESSES
Models, Methods, Directions, and Open Problems

Sean P. Meyn
Department of Electrical and Computer Engineering
and the Coordinated Sciences Laboratory
University of Illinois at Urbana-Champaign
Urbana, IL 61801, U.S.A.

May 28, 2000

Chapter 1

Stability, Performance Evaluation, and Optimization

Abstract

In this chapter we discuss various aspects of stability and control for MDP models. These ideas are centered around fluid models as approximations to the MDP model, and stochastic Lyapunov functions for verifying stability and bounding performance.

1.1 Introduction

This chapter considers stability of controlled Markov chains, and the relationship between criteria for stability; current approaches to bounding performance for the chain; and the construction of optimal policies. With stability taken as a central issue, the reader must soon realize that we will be considering large state spaces. In this chapter very little structure will be imposed on $\mathbb{X}$, or its associated sigma field $\mathcal{F}$. In other ways however our viewpoint will be fairly narrow. This chapter only considers the cost *minimization* problem, and only the average cost optimality criterion. Moreover, the assumptions that we impose will 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. An assumption of ψ-irreducibility, to be described and developed below, allows one to lift most results 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.

This surprising fact, that one does not lose much in moving from countable state space to ψ-irreducible chains on a general state space, can be regarded as a consequence of the Nummelin-Athreya-Ney construction of an 'atom'. This construction involves a time homogeneous Markov chain, where the atom $\boldsymbol{\theta}$ is a single point in an enlarged state space. The ψ irreducibility assumption makes this construction possible, and moreover ensures that the atom is reachable with positive probability from any initial condition. In this setting it is clear how to generalize many of the well known concepts in a finite state space setting. An invariant measure will be given by

$$\mu\{A\} = \mathbb{E}_\theta\Big[\sum_{t=1}^{\tau_\theta} \mathbf{1}_A(x_t)\Big], \qquad A \in \mathcal{F},$$

where $\mathbf{1}_A$ is the indicator function of the set A. In words, the quantity $\mu\{A\}$ expresses the mean number of times that the chain visits the set A before returning to $\boldsymbol{\theta}$. This expression assumes that the return time 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 above for the invariant probability μ is appealing, but it will be more convenient to work within an operator-theoretic framework. To motivate this, suppose first that we remain in the previous setting with an uncontrolled Markov chain, and suppose that do have an atom 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 $\nu(A) = P(\boldsymbol{\theta}, A)$, $A \in \mathcal{F}$, and define the 'outer product' of s and ν by

$$s \otimes \nu\,(x, A) \triangleq s(x)\nu(A).$$

We can then write in operator theoretic notation,

$$\mathbb{P}_\theta\{\tau_\theta \ge n, x_n \in A\} = \nu(P - s \otimes \nu)^{n-1}\mathbf{1}_A, \qquad n \ge 1.$$

For example, in the finite state space case the measure ν can be interpreted as a row vector, the functions s and $\mathbf{1}_A$ as column vectors, the kernel $s \otimes \nu$ is the standard (outer) product of these two vectors, and any kernel such as P or $s \otimes \nu$ can be interpreted as an $N \times N$ matrix, where N is the number of states.

Hence the invariant measure μ can be expressed in this notation as

$$\mu(A) = \sum_{n=1}^{\infty} \nu(P - s \otimes \nu)^{n-1}\mathbf{1}_A, \qquad A \in \mathcal{F}. \tag{1.1}$$

e:inv

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? First, we require a function $s\colon \mathbb{X} \to \mathbb{R}_+$ and a measure ν on $\mathcal{F}$ satisfying the *minorization condition*,

$$P(x,A) \ge s(x)\nu(A), \qquad x \in \mathbb{X},\ A \in \mathcal{F}.$$

In operator theoretic notation this is written $P \ge s\otimes\nu$, 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. For this bound to be useful we usually require s to be strictly positive on some suitably "large" set, and we require $\nu(\mathbb{X}) > 0$. Unfortunately, this 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

$$K(x,A) = (1-\beta)\sum_{t=0}^{\infty} \beta^t P^t\,(x,A), \qquad x \in \mathbb{X},\ A \in \mathcal{F}, \tag{1.2}$$

e:K

where $\beta \in]0,1[$ is some fixed constant. For a ψ-irreducible chain the required minorization always holds for the resolvent K. Much of the analysis then will involve the *potential kernel*, defined via

$$H(x,A) \triangleq \sum_{t=0}^{\infty} (K - s\otimes\nu)^t(x,A), \qquad x \in \mathbb{X},\ A \in \mathcal{F}. \tag{1.3}$$

e:potentialA

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.

The next question is stability. The ψ-irreducibility assumption will imply a reachability condition on s, analogous to the irreducibility condition that $\mathbb{P}_x\{\tau_\theta < \infty\} > 0$ for all $x \in \mathbb{X}$. Generalizations of the more useful stability conditions such as L_p bounds on τ_θ are less straightforward. We will summarize known results in Section 1.2. A key observation is that 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 equations (or ACOE). This connection will be exploited in our development of dynamic optimization theory below.

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]. 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 the stability theory is applied to the analysis of the value iteration and policy iteration algorithms. Sections 1.5 and 1.6 illustrate the theory with a detailed application to linear models, and to network scheduling. The chapter is concluded with a discussion of some open problems.

1.2 Stability

s:lyapunov

In this section we consider a Markov chain $\boldsymbol{x}$ with uncontrolled transition function P. The state space $\mathbb{X}$ is assumed to be a locally compact, separable metric space, and we let $\mathcal{F}$ 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 $\mathcal{F}$. The chain is called *ψ-irreducible* if the resolvent kernel defined in (1.2) satisfies

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

We then call ψ a *(maximal) irreducibility measure*. We let $\mathcal{F}^+$ denote the set of all measurable $h\colon \mathbb{X} \to \mathbb{R}_+$ satisfying $\psi(h) > 0$. For $A \in \mathcal{F}$ we write $A \in \mathcal{F}^+$ provided $\mathbf{1}_A \in \mathcal{F}^+$. That is, $\psi(A) > 0$. If the chain is ψ-irreducible, then from any initial condition x, the process has a chance of entering any set in $\mathcal{F}^+$ in the sense that $\mathbb{P}_x\{\tau_A < \infty\} > 0$, where τ_A is the first return time,

$$\tau_A = \min\{t \geq 1 : x_t \in A\}.$$

A function $s\colon \mathbb{X} \to \mathbb{R}_+$ and a non-trivial measure ν on $\mathcal{F}$ are called *petite* if the resolvent K satisfies the minorization condition,

$$K(x,A) \geq s(x)\nu(A), \qquad x \in \mathbb{X},\ A \in \mathcal{F}. \tag{1.4}$$

e:accessible

One can show that a Markov chain is ψ-irreducible *if and only if* there is a function $s\colon \mathbb{X} \to (0,1)$ and a probability measure ν satisfying (1.4). This bound is the most powerful consequence of the ψ-irreducibility assumption since it allows the construction of the potential kernel (1.3).

A set $C \subseteq \mathbb{X}$ is called *petite* if its indicator function $\mathbf{1}_C$ is a petite function. Equivalently, for some probability measure ν and $\delta > 0$,

$$K(x, A) \geq \delta\nu(A), \qquad x \in C,\ A \in \mathcal{F}.$$

The superlevel set $\{x : s(x) \geq \eta\}$ is always petite, for any $\eta > 0$, if the function s is petite. Consequently, for a ψ-irreducible chain, there always exists a countable covering of the state space by petite sets. The "petite property" can also be defined using hitting times: It is not difficult to show that, for a ψ-irreducible chain, the set C is petite if for each $A \in \mathcal{F}^+$, there exists $n \geq 1$, and $\delta > 0$ such that

$$\mathbb{P}_x(\tau_A \leq n) \geq \delta \qquad \text{for any } x \in C. \tag{1.5}$$

e:PetiteTime

In applications we typically find that the function s used in (1.4) can be taken positive everywhere, and *continuous.* In this case we find that every compact set is petite, in which case the Markov chain is called a *T-chain.* It will be convenient to restrict attention to T-chains in Section 1.4.

1.2.2 Recurrence and stability

The fundamental stability assumption for a Markov chain is the property that the state visit 'important' sets with probability one from any starting point. A ψ-irreducible chain is called *recurrent* if there exists a set $\mathbb{X}_0 \in \mathcal{F}$ satisfying $\psi\{\mathbb{X}_0^c\} = 0$, and $\mathbb{P}_x\{\tau_A < \infty\} = 1$ for any $A \in \mathcal{F}^+$, and any $x \in \mathbb{X}_0$. If in addition the chain admits an invariant probability measure μ, then the chain is called *positive recurrent.* Although these definitions require us to consider an infinite number of initial states and sets A, there are several equivalent characterizations of recurrence which are easier to verify.

t:recurrence

Theorem 1 *The following are equivalent for a ψ-irreducible Markov chain $\boldsymbol{x}$:*

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

(ii) *For some petite set C,*

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

(iii) *For any pair (s, ν) satisfying the minorization condition (1.4) with $s \in \mathcal{F}^+$,*

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

If any of these three 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

$$\mu^*(A) = \nu H\{A\} = \sum_{t=0}^{\infty} \nu(K - s\otimes\nu)^t\{A\}, \qquad A \in \mathcal{F}. \tag{1.6}$$

e:inv2

Proof. The proof can be found in [44, 39]. However it will be useful to explain the main ideas, and in particular show why μ^* defines an invariant measure.

To show why (iii) should hold for a recurrent chain we first make some definitions. 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 $H_1 = H$ is the potential kernel. These kernels are uniformly bounded in x for any $\lambda > 1$. We denote $\alpha_\lambda = \nu(H_\lambda s)$ and $\beta_\lambda = \nu(G_\lambda s)$.

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

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

Iterating this equation we find that 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 $\lambda > 1$. Since all of the terms are positive we also see that $\alpha_\lambda < 1$. Letting $\lambda \downarrow 1$ and applying the monotone convergence theorem we do see that $\alpha_1 \le 1$ and in fact that,

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

from which we see that $\nu H s = \alpha_1 = 1$ if and only if $\beta_1 = \infty$. It remains to show that an infinite value for β_1 is equivalent to recurrence. This is not difficult to believe, but due to lack of space we ask that the reader look elsewhere, e.g. [44, 39].

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

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

Now by recurrence and (iii) we have $\mu^*(s) = 1$, which shows that μ^* is K-invariant: $\mu^* K = \mu^*$. Using the identity

$$PK = KP = \beta^{-1}K + (1 - \beta^{-1})I, \tag{1.7}$$

e:resolvEqn

one can conclude that μ^* must then be P-invariant. ∎

The invariant measure given in (1.6) will be finite, so the chain $\boldsymbol{x}$ is positive recurrent, provided that the *mean* return time to a petite set C is bounded:

$$\sup_{x \in C} \mathbb{E}_x[\tau_C] < \infty. \tag{1.8}$$

e:mean

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.

While these definitions lead to an elegant theory, in practice one can typically take $\mathbb{X}_0 = \mathbb{X}$ in the definition of recurrence. 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 c-Regularity and Lyapunov functions

The next level of stability that we consider is more closely 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. For a ψ-irreducible chain, a set $S \in \mathcal{F}$ is called *c-regular* if for any $A \in \mathcal{F}^+$,

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

From the characterization in (1.5) we see that a c-regular set is always petite. 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 by using the previous representation of the invariant measure one can show that a c-regular chain possesses an invariant probability measure μ satisfying $\mu(c) \stackrel{\Delta}{=} \int c(x)\,\mu(dx) < \infty$. MT The following result is a consequence of the f-Norm Ergodic Theorem of [39, Theorem 14.0.1].

t:little-f-norm **Theorem 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:

$$\begin{array}{lrcll} \textbf{\textit{(i)}} & \displaystyle\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} g(x_t) & = & \mu(g), & a.s.\ [\mathbb{P}_x]. \\ \textbf{\textit{(ii)}} & \displaystyle\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_x[g(x_t)] & = & \mu(g). & \end{array}$$

■

Thus we see that one level of complexity is removed when considering c-regular chains since the steady state performance as captured by $\mu(g)$ does not depend upon the initial condition of the chain.

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

$$PV\,(x) \stackrel{\Delta}{=} \mathbb{E}[V(x_{t+1}) \mid x_t = x] \le V(x) - c(x) + \bar{J}, \qquad x \in \mathbb{X}. \tag{1.9}$$ e:cFoster

However, for this to imply any form of stability, the difference $c(x) - \bar{J}$ must be positive for 'large' x. This requires some assumptions on c. We say that c is *near-monotone* if the sublevel set $S_\eta \stackrel{\Delta}{=} \{x \in \mathbb{X} : c(x) \le \eta\}$ is petite for any $\eta < \|c\|_\infty$. The supremum norm $\|c\|_\infty = \sup_{x\in\mathbb{X}} c(x)$ may be infinite. The function is called *norm-like* if the sublevel set S_η is a precompact subset of the metric space $\mathbb{X}$ for *any* η. Related assumptions on c are used in araborferghomar93,bor91,MT [1, 5, 39].

t:LyapunovRegular **Theorem 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 (1.9), then there exists $d_0 < \infty$ such that for each $A \in \mathcal{F}^+$,*

$$\mathbb{E}_x\Big[\sum_{t=0}^{\tau_A} c(x_t)\Big] \le d_0 V(x) + d(A), \qquad x \in \mathbb{X}, \tag{1.10}$$

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

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

$$V^*(x) = \mathbb{E}_x\Big[\sum_{t=0}^{\tau_S} c(x_t)\Big], \qquad x \in \mathbb{X}, \tag{1.11}$$

is a near-monotone solution to (1.9).

Proof. The bound (1.9) is equivalent to the drift condition $PV_0 \le V_0 - c + b\mathbf{1}_S$, where S is petite: if (1.9) 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

1.3.1 Poisson's equation

Poisson's 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 then define the generator as $\Delta = P - I$, and the Poisson equation then takes on the exact same form, where we take $f = c - \mu(c)$.

The probabilistic motivation for looking at this equation follows from our prior stability analysis. First note that the drift inequality (1.9) suggests a simple approach to obtaining performance bounds. By iterating this equation one obtains,

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

Dividing by n and letting $n \to \infty$ then gives the upper bound,

$$J(c) \triangleq \limsup_{n\to\infty} \frac{1}{n} \mathbb{E}_x\Bigl[\sum_{t=0}^{n-1} c(x_t)\Bigr] \le \bar{J}.$$

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 evidentally $\mu(c)$, with μ equal to the invariant probability for the chain. The following result surveys the relevant consequences of c-regulariry, and introduces the form of the Poisson equation that we will analyze in the remainder of this chapter.

t:fish **Theorem 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 the Poisson equation*

$$Ph\,(x) \triangleq \mathbb{E}_x[h(x_{t+1}) \mid x_t = x] = h(x) - c(x) + J, \qquad x \in \mathbb{X}, \quad (1.12)$$ e:fish

where $J = \mu(c)$.

(ii) *One solution to* e:fish *(1.12) may be expressed,*

$$h(x) = KH\bar{c}\,(x) = \sum_{t=1}^{\infty} (K - s\otimes\nu)^t \bar{c}\,(x), \qquad x \in \mathbb{X}, \qquad (1.13)$$ e:fishFormula

where $\bar{c} = c - J$, and the pair (s,ν) is petite with $s \in \mathcal{F}^+$.

(iii) *Suppose moreover that the function c is near-monotone. Then the solution h to* e:fishFormula *(1.13) 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 $J = \pi(c)$, then there exists a constant a such that

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

(iv) *If V is any solution to* e:cFoster *(1.9) with $\bar{J} < \|c\|_\infty$ and c near-monotone, then the solution* e:fishFormula *(1.13) satisfies the uniform upper bound, for some $d_0 < \infty$,*

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

Proof. Again, the proof can be found elsewhere (see [42], following [45, 22]). However, the fact that the solution to the Poisson equation (1.12) can be taken as in (1.13) is easy to explain, and shows some attractive symmetry with the earlier construction of an invariant measure. Consider first the function

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

Observe that by the definition of J and $\bar{c}$, and the construction of μ^*, we must have $\nu(h_0) = 0$. Thus,

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

That is, h_0 solves the Poisson equation for the kernel K. By again applying the identity (1.7) we see that $h = Kh_0 = (K - s \otimes \nu)h_0$ solves the Poisson equation for original transition kernel P. ∎

1.3.2 Simulation

We have now seen that the Poisson equation has a direct role in performance evaluation. Although we have not given any explicit algorithms, it is clear that an approximation of the solution h will lead to an approximation of $J = \mu(c)$. 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

$$\gamma_c^2 = \mu(h^2) - \mu((Ph)^2) = 2\mu(h\bar{c}) - \mu(\bar{c}^2), \tag{1.14}$$ e:cltVar

with h any solution to (1.12) [39].

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 (1.12): 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].

This is a useful result for our purposes since we will discover such approximations when we attempt to solve some optimization problems below.

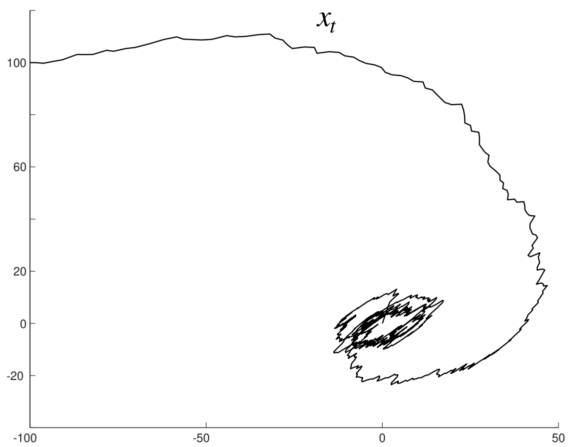
1.3.3 Examples

In this chapter we will restrict ourselves to 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 two sections in this chapter.

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

$$x_{t+1} = Ax_t + w_{t+1}, \qquad t \in \mathbb{N}, \tag{1.15}$$ e:linear-0

where $x_t, w_t \in \mathbb{R}^d$, and $\boldsymbol{w}$ is i.i.d. with $w_t \sim N(0, \Sigma)$. Let F by any matrix of suitable dimension satisfying $FF^T = \Sigma$. If F is $d \times q$ for some q, then the *controllability matrix* is the $d \times (dq)$ matrix $\mathcal{C} \triangleq [A^{d-1}F | A^{d-2}F | \cdots | AF | F]$, and the pair (A, F) is called *controllable* if the matrix $\mathcal{C}$ has rank d. The process is ψ-irreducible with ψ equal to Lebesgue measure if the pair (A, F) is controllable, since in this case $P^t(x, \cdot\,)$ is equivalent to Lebesgue measure for any x, and any $t \geq d$. By continuity of the model it is easy to check that (1.4) holds with s continuous, and ν equal to normalized Lebesgue measure on an open ball in $\mathbb{R}^d$. We conclude that all compact sets are petite if the controllability condition holds, so that $\boldsymbol{x}$ is a T-chain. A sample path from a particular two dimensional linear model is shown in Figure 1.1 where it can be seen that when the state is large, the sample path behavior appears almost deterministic.



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

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

$$A^T P A = P - Q. \tag{1.16}$$ e:lyapEqn

If $P \geq 0$, then $V(x) = \frac{1}{2}x^T P x$ is a solution to (1.9). In fact, the function V is also the essentially unique solution to Poisson's equation, with $J = \frac{1}{2}\mathrm{trace}\,(P\Sigma)$.

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 (1.16) 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.

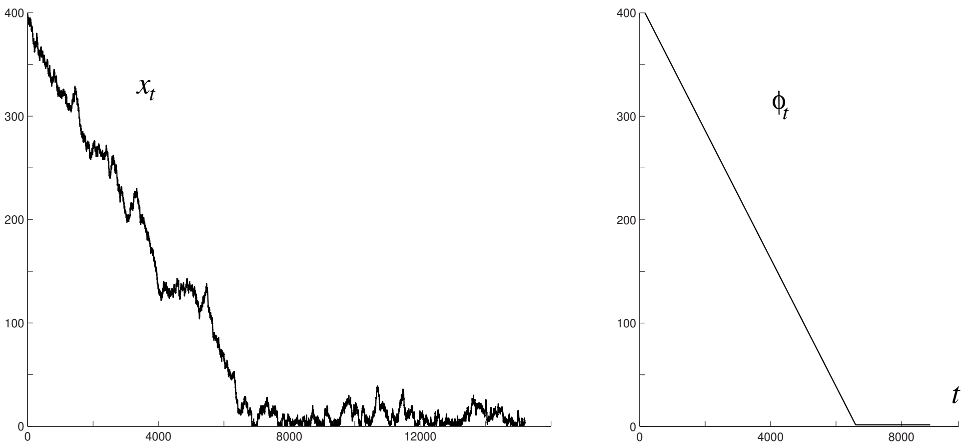
When the arrival stream is renewal, and the service times are i.i.d., then the waiting time for a simple queue can be modeled as a Markov chain with state space $\mathbb{X} = \mathbb{R}_+$. The dynamics take the form of a one dimensional linear state space model, where the state space is constrained 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 Bernouli, 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 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 $c(x) = x$. To construct a Lyapunov function we would ideally like to compute the expected sum given in (1.11), 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 $\dot{\phi} = (-\mu + \lambda)\pi(\phi)$, where π is again equal



f:mm1 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 $\dot{\phi} = (-\mu + \lambda)\pi(\phi)$ starting from the same initial condition.

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 $V^*(x) \approx$

$$\begin{aligned} V(x) &\triangleq \int_0^\infty \phi(t)\, dt, \quad \phi(0) = x, \\ &= \tfrac{1}{2}\frac{x^2}{\mu - \lambda}. \end{aligned} \tag{1.17}$$ e:fluidValue

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 (1.9) does hold with this function V derived from the fluid model, where e:cFoster

$\bar{J} = (2(\mu - \lambda))^{-1}$, 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 $c(x) = x$, the Poisson equation for the M/M/1 queue becomes

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

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 (1.17).

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.

The M/M/1 queue illustrates nicely the difficulties one faces in using simulation to esimate steady state performance measures since the time-average variance constant γ_c^2 grows quickly with the 'system load'. To obtain bounds on γ_c^2 note firstly that the solution to Poisson's equation for the M/M/1 queue is a quadratic of the form $h(x) = b(1-\rho)^{-1}(x^2 + x)$, where b is a bounded function of λ. Using (1.14) we conclude that γ_c^2 is of order $(1-\rho)^{-4}$ in this example since pth moments for the M/M/1 queue are of order $(1-\rho)^{-p}$.

It is shown in [25] that this growth rate on γ_c^2 continues to hold for a general class of network models. The large variance indicates that in heavy traffic (where $\rho \sim 1$) it is computationally expensive to compute the mean performance through standard Monte Carlo simulation. This provides ample motivation to find methods for approximating solution's to Poisson's equation to construct reduced variance estimators for network models. The use of the fluid value function is one promising approach [25].

1.4 Optimization

s:opt

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

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 $\mathcal{F}$ 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 the 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 $\{P_a(x, B)\}$ which describe the probability that the next state is in B for any $B \in \mathcal{F}$ given that the current state is $x \in \mathbb{X}$, and the current action chosen is $a \in \mathbb{A}$. These are assumed to be probability measures on $\mathcal{F}$ for each state-action pair (x, a), and measurable functions of (x, a) for each $B \in \mathcal{F}$. We will primarily restrict attention to stationary Markov policies. Recall that such a policy, denoted $\pi \in \Pi^S$, is a measurable function $\pi : \mathbb{X} \to \mathbb{A}$ such that $\pi(x) \in \mathbb{A}(x)$ for all x. When the policy π is applied to the MDP, then the action $\pi(x)$ is applied whenever the MDP is in state x, independent of the past and independent of the time-period. We shall write $P_\pi(x, B) = P_{\pi(x)}(x, B)$ for the transition law corresponding to a policy π.

The state process $\boldsymbol{x}^\pi \triangleq \{x_t^\pi : t \geq 0\}$ of the MDP is, for each fixed policy π, a Markov chain on $(\mathbb{X}, \mathcal{F})$, and we write the t-step transition probabilities for this chain as

$$P_\pi^t(x, B) = \mathbb{P}(x_t^\pi \in B \mid x_0^\pi = x), \qquad x \in \mathbb{X}, \quad B \in \mathcal{F}, \quad t \in \mathbb{N}.$$

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

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

In the remander of this section we will first recall the ACOE for an MDP, and then develop two algorithms which can be used to construct solutions to these equations. The value iteration algorithm (VIA) will be considered first, which is simply the successive approximation procedure. The results describe mainly stability of the policies generated by the algorithm, and some bounds on steady state performance. These results are based on [10]. The policy iteration algorithm (PIA), first proposed in [31], may be viewed as a version of the Newton Raphson method, and is consequently considerably more complex than value iteration. It is however far easier to analyze.

We find under the assumptions we impose that it exhibits nearly monotone convergence, and that the algorithm does generate successively better policies. These results are taken from [42].

More background and convergence results for such algorithms may be found in [1, 5, 7, 8, 27, 29, 28, 48, 54].

1.4.1 Regular policies and the ACOE

We now suppose that a one-step cost function $c\colon \mathbb{X}\times\mathbb{A}\to[1,\infty)$ is given: we assume below that c satisfies a near-monotone condition so that the results of Section 1.2 and 1.3 may be applied. For any policy π we denote $c_\pi(y)=c(y,\pi(y))$, and we denote the steady state average cost by

$$J(\pi,x)\triangleq\limsup_{n\to\infty}\frac{1}{n}\,\mathbb{E}_x\Bigl[\sum_{t=0}^{n-1}c_\pi(x_t^\pi)\Bigr].$$

A policy π_* will be called *optimal* if $J(\pi_*,x)\le J(\pi,x)$ for all policies π, and any initial state x.

The policy π is called *regular* if the controlled chain $\boldsymbol{x}^\pi$ is a c_π-regular Markov chain. This is a natural and highly desirable stability property for the controlled process. If the policy is regular, then necessarily an invariant probability measure μ_π exists such that $\mu_\pi(c_\pi)<\infty$. Moreover, for a regular policy, the resulting average cost is independent of the starting point of the chain: $J(\pi,x)\equiv\mu_\pi(c_\pi)$.

When $J_*=J(\pi_*,x)$ is independent of x, then the associated ACOE are given as follows, where the function h_* is known as the *relative value function*.

$$\begin{aligned} J_*+h_*(x) &= \min_{a\in\mathbb{A}(x)}[c(x,a)+P_a h_*(x)] && (1.18)\ \boxed{\texttt{e:OE1}}\\ \pi_*(x) &= \arg\min_{a\in\mathbb{A}(x)}[c(x,a)+P_a h_*(x)], \qquad x\in\mathbb{X}. && (1.19)\ \boxed{\texttt{e:OE2}} \end{aligned}$$

If a policy π_*, a measurable function h_*, and a constant J_* exist which solve these equations, then typically the policy π_* is optimal (see for example [1, 5, 26, 48, 53] for a proof of this and related results).

t:OE **Theorem 5** *Suppose that the following conditions hold*

(a) *The pair (J_*,h_*) solve the optimality equation (1.18);*

(b) *The policy π_* satisfies (1.19), so that*

$$c_{\pi_*}(x) + P_{\pi_*} h_*(x) \le c(x,a) + P_a h_*(x), \qquad x \in \mathbb{X},\ a \in \mathbb{A}(x).$$

(c) *For any $x \in \mathbb{X}$, and any policy π satisfying $J(\pi, x) < \infty$,*

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

Then π_* is an 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 $J(\pi, x) \ge J_*$ for all policies π, and all initial states x. ∎

The assumption (c) is unfortunate, but examples show that some additional conditions on h_* are required (see e.g. page 87 of [5], Chapter 7 of [53], or the examples in [1, 16, 48, 50]). The following result gives a condition implying (c) which is often verifiable in practice, as we shall see in Section 1.5 and Section 1.6.

If the controlled chain $\boldsymbol{x}^\pi$ is ψ_π-irreducible and the resulting cost $J_\pi \triangleq \mu_\pi(c_\pi) = \int c_\pi(x)\mu_\pi(dx)$ is finite, let S_π denote any fixed c_π-regular set for which $\mu_\pi(S_\pi) > 0$. We then define the function

$$V_\pi(x) = \mathbb{E}_x\Bigl[\sum_{t=0}^{\tau_\pi - 1} c_\pi(x_t^\pi)\Bigr], \tag{1.20}$$ e:optV

where $\tau_\pi = \tau_{S_\pi}$. Since $\mu_\pi(S_\pi) > 0$, the function V_π is a.e. $[\mu_\pi]$ finite-valued [39, Theorem 14.2.5]. Note that by [39, Theorem 14.2.3], the particular c_π-regular set S_π chosen is not important. If S_π^1 and S_π^2 give rise to functions V_π^1 and V_π^2 of the form (1.20), then for some constant $\gamma \ge 1$,

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

t:comparable **Theorem 6 (Meyn [42])** *Suppose that*

(a) *The optimality equations (1.18,1.19) hold for (π_*, h_*, J_*), with h_* bounded from below;*

(b) *For any policy π the average cost $K_\pi c_\pi$ is finite and norm-like, and all compact sets are petite for the Markov chain $\boldsymbol{x}^\pi$;*

(c) *For any policy π there exists some constant $d_0 = d_0(w) < \infty$ such that,*

$$|h_*(x)| \le d_0 V_\pi(x), \qquad x \in \mathbb{X}. \tag{1.21}$$ e:relativeBound

Then π_* is a regular, optimal policy.

Proof. To show that π_* is regular we consider the Poisson equation for the Markov chain with transition kernel K_{π_*}:

$$K_{\pi_*} h_* = h_* - K_{\pi_*} c_{\pi_*} + J_*.$$

Under the assumption that h_* is bounded from below, it follows from Theorem 2 that the "K_{π_*}-chain" is $K_{\pi_*} c_{\pi_*}$-regular, and regularity of π_* follows.

To prove optimality suppose that π is any policy, and note first that if $J(\pi, x) = \infty$ for all x, then there is nothing to prove. If not, then since all compact sets are petite, the Markov chain $\boldsymbol{x}^\pi$ is a positive recurrent T-chain, with unique invariant probability μ_π, and $J_\pi \triangleq \mu_\pi(c_\pi)$ is finite [39, Theorem 14.0.1].

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

$$P_\pi^n V_\pi = V_\pi - \sum_{t=0}^{n-1} P_\pi^t (c_\pi - s_\pi),$$

where $s_\pi \geq 0$ and satisfies $\mu_\pi(s_\pi) = \mu_\pi(c_\pi)$. This function can be written explicitly as

$$s_\pi(x) = \int_{S_\pi} P_\pi(x, dy)\, \mathbb{E}_y \Big[\sum_{t=0}^{\tau_\pi - 1} c_\pi(x_t) \Big]$$

It follows from [39, Theorem 14.0.1] that $P_\pi^n V_\pi(x)/n \to 0$ as $n \to \infty$ for a.e. $[\mu_\pi]$ $x \in \mathbb{X}$. By (1.21), we also have $P_\pi^n h_*(x)/n \to 0$, and it easily follows that $J(\pi, x) \geq J(\pi_*)$ for such x. To generalize to arbitrary x we use the norm-like assumption on c which implies that $J(\pi, x) \geq \mu_\pi(c_\pi)$ for all $x \in \mathbb{X}$ (for details see [42]). ∎

1.4.2 Value iteration

s:via

The ACOE 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 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

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

which then makes it possible to compute the next policy π_{n+1}.

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

$$V_n(x) = \min \mathbb{E}_x\Bigl[\sum_{t=0}^{n-1} c(\Phi(t), a(t)) + V_0(\Phi(n))\Bigr], \tag{1.22}$$

e:ValueFunction

where $\{a(t) : t \geq 0\}$ is a sequence of actions determined by some Markov policy, and the minimum in (1.22) is with respect to all such 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

$$K_n \triangleq (1-\beta)\sum_{t=0}^{\infty} \beta^t P_n^t, \qquad n \geq 0, \tag{1.23}$$

e:Kn-def

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 finite measure on $\mathcal{F}$ - we will impose conditions below which ensure that $\nu(V_n)$ is finite-valued. We then define for each n the normalized value function, and the incremental cost,

$$h_n(x) = V_n(x) - \nu(V_n); \qquad \gamma_n(x) = V_{n+1}(x) - V_n(x), \qquad x \in \mathbb{X}, n \in \mathbb{N}. \tag{1.24}$$

e:hg-def

From the definitions, for each n we have the familiar looking identity $P_n h_n = h_n - c_n + \gamma_n$. For this to imply any form of stability for the policy π_n we require that γ_n be bounded from above. When this is the case we define $\bar{J}_n = \sup_x \gamma_n(x)$, so that we have the following solution to (1.9):

$$P_n V_n \leq V_n - c_n + \bar{J}_n. \tag{1.25}$$

e:almostFish-a

Useful bounds on γ_n will in general require conditions on the initial condition V_0. We show here that a sufficient condition is that V_0 act as a Lyapunov function for *some* policy. We assume that at least one regular policy π_{-1} exists, and that for some constant $\bar{J} < \infty$,

$$P_{\pi_{-1}} V_0 \leq V_0 - c_{\pi_{-1}} + \bar{J}. \tag{1.26}$$

e:stable-ic

The existence of a pair (V_0, π_{-1}) satisfying (1.26) is a natural stabilizability assumption on the model.

The assumptions below ensure that the algorithm can be initialized to generate stabilizing policies. Condition (A2) relates the average optimality problem with the discounted optimal control problem: this assumption is satisfied if the state dependent cost V_β is norm-like, where $V_\beta(x)$ denotes the optimal cost for the discounted optimal control problem when the discount factor is equal to β, and the initial condition is x. Assumption (A3) is a uniform minorization condition on the MDP model.

(A1) For each n, if the VIA yields a value function $V_n\colon \mathbb{X} \to \mathbb{R}_+$, then the minimization

$$\pi_n(x) \triangleq \mathop{\arg\min}_{a \in \mathbb{A}(x)} \bigl[c(x,a) + P_a V_n\,(x)\bigr]$$

admits a measurable solution π_n.

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

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

(A3) There is a fixed probability ν on $\mathcal{F}$, 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 (A1),

$$K_n(x,A) \geq \delta\nu(A) \qquad \text{for all } x \in S,\ A \in \mathcal{F}, \tag{1.27}$$ e:nuAccessibilityVIA

where S denotes the precompact set

$$S = \{x : \underline{c}(x) \leq 2\bar{J}\}. \tag{1.28}$$ e:S-def-VIA

The following result is largely taken from chemey99a [10].

t:VIA-c-regular **Theorem 7** *Suppose that (A1) and (A2) hold, and suppose that for each of the policies $\{\pi_n\}$ obtained through the VIA, all compact sets are petite for the Markov chain $\boldsymbol{x}^{\pi_n}$. Assume moreover that the initialization V_0 satisfies* e:stable-ic *(1.26). 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) *If in addition Assumption (A3) holds, then the sequence $\{h_n\}$ is uniformly bounded from below.*

Proof. The main idea is to apply the following bound on the resolvent, which follows from (1.25):

$$K_n V_n \le V_n - \frac{\beta}{1-\beta} K_n c_n + \frac{\beta}{1-\beta} \bar{J}_n. \tag{1.29}$$

This inequality is a version of (1.9) since $V_n \ge 0$, provided that $\bar{J}_n$ is finite.

The minimization in the value iteration algorithm immediately leads to the bound $P_n \gamma_n \ge \gamma_{n+1}$. From this we deduce by induction that the $\bar{J}_n$ are finite and decreasing, which proves (ii). The initialization of the induction relies on the assumption that the initial condition V_0 satisfies (1.26). Applying Theorem 3, which is applicable under (A2) 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 $h_n(x) \ge -\nu(V_n) > -\infty$ for all x. It remains to obtain a bound independent of n. For any n we have

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

Letting $s = \delta \mathbf{1}_S$ we then obtain, for some measure ν,

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

and by iteration, for any N,

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

By c_n-regularity of the nth chain one can show that for any x,

$$\sum_{i=0}^{\infty} (K_n - s\otimes\nu)^i c\,(x) < \infty.$$

Since $c \ge 1$ it follows that $(K_n - s\otimes\nu)^N \mathbf{1}\,(x) \to 0$ as $N \to \infty$ for any x. This then gives the bound

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

The proof is completed on noting that $\sum_{i=0}^{\infty} (K_n - s\otimes\nu)^i s\,(x) \le (\delta(1-\delta))^{-1}$ for all x (this can be seen by noting that this sum is proportional to the hitting probability to an atom for the split chain as in [44, 39]). ∎

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 (A1), (A2), and a strengthening of (A3) 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 [28], or the stronger stability assumption imposed in [15, 55].

1.4.3 Policy iteration

The PIA, which is again a recursive algorithm for generating useful policies, follows naturally as a refinement of the VIA. Recall that the key observation in Section 1.4.2 was the drift inequality,

$$P_{n-1}V_{n-1} \le V_{n-1} - c_{n-1} + \bar{J}_{n-1},$$

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 the drift inequality (1.9), 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 $\bar{J}$. If the function V_{n-1} is replaced by the solution to Poisson's equation in the VIA recursion then 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 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 policy π_n by choosing, for each x,

$$\pi_n(x) = \mathop{\arg\min}_{a \in \mathbb{A}(x)} \bigl[c(x,a) + P_a h_{n-1}(x)\bigr]. \tag{1.30}$$

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

Recall that the relative value functions $\{h_n\}$ are not uniquely defined: If h_n satisfies Poisson's equation, then so does $h_n + b$ for any constant b.

The main results to follow all apply to specific normalized versions of the relative value functions.

We find that the PIA is more easily analysed then the VIA since we know a great deal about the structure of solutions to Poisson's equation. To begin, consider the pair of equations

$$P_n h_n = h_n - \bar{c}_n; \tag{1.31}$$

$$P_n h_{n-1} = h_{n-1} - \bar{c}_n + \gamma_n, \tag{1.32}$$

where $\bar{c}_n = c_n - J_n$, and γ_n is *defined* through (1.32). From the minimization (1.30) we have

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

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 $\gamma_n(x) \leq J_{n-1} - J_n$, $x \in \mathbb{X}$, which shows that the PIA automatically generates solutions to (1.9), provided the functions $\{h_n\}$ can be taken to be positive. Through these observations we will show that the sequence $\{J_n\}$ converges monotonically, which shows that the limit superior of $\{\gamma_n\}$ is bounded from above by zero. Under suitable conditions we also show that the sequence $\{h_n\}$ is bounded from below, and this then gives a lower bound on $\{\gamma_n\}$: Since $\mu_{\pi_n}(c_n) = J_n$, it follows from the Comparison Theorem [39, p. 337] that $\mu_{\pi_n}(\gamma_n) \geq 0$.

Thus, for large n, the error term γ_n is small, and hence the function h_{n-1} *almost* solves the Poisson equation for P_n. One might then expect that h_n will be close to h_{n-1}. Under mild conditions, this is shown to be true in a very strong sense.

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 (1.23). To invoke the algorithm we must again ensure that the required minimum exists. Condition (A2) is also identical to our earlier assumption on the one step cost in Section 1.4.2.

(A1) For each n, if the PIA yields a triplet $(\pi_{n-1}, h_{n-1}, J_{n-1})$ 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 the minimization

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

admits a measurable solution π_n.

(A2) For each fixed x, the function $c(x,\cdot)$ is norm-like on $\mathbb{A}$, and 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,

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

Under Assumptions (A1) and (A2), the algorithm produces stabilizing policies recursively. The proof is identical to Theorem t:VIA-c-regular 7.

t:PIA-c-regular

Theorem 8 (Meyn mey97b [42]) *Suppose that (A1) and (A2) hold, and that for some n, the policies $\{\pi_i : i < n\}$ and relative value functions $\{h_i : i < n\}$ are defined through the policy iteration algorithm. Suppose moreover that*

(a) *The relative value function h_i is bounded from below, $i \leq n-1$.*

(b) *All compact sets are petite for the Markov chains $\{\boldsymbol{x}^{\pi_i},\ i \leq n-1\}$, and for $\boldsymbol{x}^{\pi_n}$, where π_n is a policy given in (A1).*

Then, the PIA admits a solution (π_n, h_n, J_n) such that

(i) *The relative value function h_n is bounded from below;*

(ii) *Each of the policies $\{\pi_i : i \leq n\}$ is regular.*

(iii) *For all $0 \leq i \leq n$ the constant J_i is the cost at the ith stage: $J_i = J(\pi_i)$, and the costs are decreasing:*

$$J_0 \geq J_1 \geq \cdots \geq J_n;$$

To obtain convergence of the algorithm, we strengthen assumption (b) of Theorem t:PIA-c-regular 8 to the following uniform accessibility condition:

(A3) There is a fixed probability ν on $\mathcal{F}$, 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 $(\pi_{n-1}, h_{n-1}, J_{n-1})$ with h_{n-1} bounded from below, then for any policy π_n given in (A1),

$$K_n(x,A) \geq \delta\nu(A) \qquad \text{for all } x \in S,\ A \in \mathcal{F}, \tag{1.33}$$ e:nuAccessibility

where S denotes the precompact set

$$S = \{x : \underline{c}(x) \leq 2J_0\}. \tag{1.34}$$ e:S-def-PIA

If the assumptions of Theorem 8 hold, so that the relative value functions are bounded from below, then the minimal solution h_n to Poisson's equation (1.12) which is bounded from below, and which satisfies $\nu(h_n) = 0$, is defined through the potential kernel in equation (1.13) of Theorem 4. These particular versions are in fact convergent:

Theorem 9 (Meyn [42]) *Suppose that (A1)-(A3) hold, and that the initial policy π_0 is regular. Then for each n the PIA admits a solution (π_n, h_n, J_n) such that π_n is regular, and the sequence of relative value functions $\{h_n\}$ defined in (1.13) satisfy,*

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

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

(ii) *There exists a sequence of functions $\{g_n : n \geq 0\}$ such that*

$$g_n(x) \leq g_{n-1}(x) \leq \cdots \leq g_0(x), \qquad x \in \mathbb{X}, \ n \geq 0,$$

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

$$g_n(x) = \alpha_n h_n(x) + \beta_n, \qquad n \geq 0, \ x \in \mathbb{X},$$

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

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

We have already remarked that the average cost optimal control problem is plagued with counterexamples. It is of some interest then to see why Theorem 9 does not fall into any of these traps. Consider first counterexamples 1 and 2 of [50, p. 142]. In each of these examples the process, for any policy, is completely non-irreducible in the sense that $\mathbb{P}(x_t^\pi < x_0^\pi) = 0$ for all times t, and all policies π. It is clear then from the cost structure that the bound (1.33) on the resolvent cannot hold. A third example is given in the Appendix of [50]. Here (1.33) 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, at least for the policies $\{\pi_n\}$. 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 counterexamples that optimal stationary policies may not exist.

Now that we know that $\{h_n\}$ is pointwise convergent to a function h, we can show that the PIA yields a solution to the ACOE. Theorem 10 is similar to Theorem 4.3 of [26] which also requires a continuity condition related to (A4). Weaker conditions are surely possible for a specific application.

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

t:PIA-opt **Theorem 10 (Meyn [42])** *Suppose that (A1)-(A3) hold, and that the initial policy π_0 is regular. Then,*

(i) *The PIA produces a sequence of solutions (π_n, h_n, J_n) such that $\{h_n\}$ is pointwise convergent to a solution h of the optimality equation (1.18), and any policy π which is a pointwise limit of π_n satisfies (1.19). Moreover, the costs $\{J_n\}$ are decreasing with n.*

(ii) *Any limiting policy π is c_π-regular, so that for any initial condition $x\in\mathbb{X}$,*

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

Observe that Theorem 10 still does not claim that the solution to (1.18) gives rise to an optimal policy. For this one must impose extra conditions. For instance, the limiting policy will be optimal provided the relative value function h satisfies the conditions of Theorem 5 or Theorem 6. Because of the monotone convergence, this will be implied by a bound on the initial condition (π_0, h_0).

We now illustrate the theory with some general examples.

1.5 Linear models

s:linearEx

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

$$x_{t+1}=Ax_t+Ba_t+w_{t+1},\qquad t\in\mathbb{N},\tag{1.35}$$ e:linear

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

$$c(x,w)=\tfrac{1}{2}x^TQx+\tfrac{1}{2}w^TRw,\tag{1.36}$$ e:lqg-cost

with $Q \geq 0$, and $R > 0$. If $\boldsymbol{w}$ is i.i.d., then this is a Markov decision process with transition function

$$P_a(x, C) = \mathbb{P}(w_1 + Ax + Ba \in C).$$

The optimization of $J(\pi, x)$ is known as the LQG (linear-quadratic-Gaussian) problem in the special case where $\boldsymbol{w}$ is *Gaussian* white noise. Note that the assumption $c \geq 1$ fails in this example. However, c is positive, so that we can add 1 to the cost function to satisfy the desired lower bound on c and the MDP is essentially unchanged.

In the Gaussian case one may obtain a solution (π_*, h_*, J_*) to the ACOE with h_* quadratic, and π_* linear in x, by solving a Riccati equation. Why then should the solution give rise to an optimal policy? This question can be answered using Theorem 6: Suppose that π is any (measurable) nonlinear feedback control. From the assumption that $\Sigma > 0$, the process $\boldsymbol{x}^\pi$ is ψ-irreducible, with ψ = Lebesgue measure. The function V_π given in (1.20) then satisfies the lower bound

$$V_\pi(x) \geq \sum_{t=0}^{\infty} 2^{-k} P_\pi^t c_\pi\,(x) \geq V_{\frac{1}{2}}(x) - b_\pi, \tag{1.37}$$

e:lqg-A4

where $V_{\frac{1}{2}}$ is the value function for the discounted problem, and b_π is a finite constant. To see this, let $\tau_\pi = \tau_{S_\pi}$, and write

$$\begin{aligned} V_{\frac{1}{2}}(x) \leq \sum_{t=0}^{\infty} 2^{-t} P_\pi^t c_\pi\,(x) \quad &= \quad \mathbb{E}_x[\sum_{t=0}^{\tau_\pi - 1} 2^{-k} c_\pi(x_t^\pi)] + \mathbb{E}_x[\sum_{t=\tau_\pi}^{\infty} 2^{-k} c_\pi(x_t^\pi)] \\ &\leq \quad V_\pi(x) + \mathbb{E}_x[2^{-\tau_\pi} \sum_{t=0}^{\infty} 2^{-k} c_\pi(x_{\tau_\pi + t}^\pi)] \end{aligned}$$

The lower bound (1.37) on V_π then follows from the strong Markov property, and regularity of the set S_π. If $(A, \sqrt{Q})$ is observable so that $V_{\frac{1}{2}}$ dominates a positive definite quadratic, then we see from Theorem 6 that the linear/quadratic solution (π_*, h_*) to the ACOE does yield an optimal policy over the class of all nonlinear feedback control laws.

The linear state space model gives an excellent test-case for interpreting the assumptions of Theorems 7–9. Condition (A1), which demands the existence of a minimizing policy, is satisfied because the model is continuous. The assumption of an initial regular policy is simply stabilizability of (A, B). The norm-like condition (A2) is implied by the standard observability condition on $(A, \sqrt{Q})$, where Q is the state weighting matrix given in (1.36) [42].

To verify the uniform accessibility condition (1.27) in (A3), suppose the initial policy π_0 is linear, so that each subsequent policy is of the form $\pi_n(x) = -K_n x$, and let $A_n = A - BK_n$ denote the closed loop system matrix. Then the accessibility condition (A3) holds if $\Sigma > 0$, and the steady state costs $\{J_n\}$ are bounded. For both the PIA and the VIA, the boundedness condition holds automatically under (A1) and (A2) through stability of the algorithm, which is guaranteed by Theorems 7 and 8.

From these results it follows that Theorem 9 recovers known properties of the Newton-Raphson technique applied to the LQG problem. Consider the well known decreasing property of the solutions $\{\Lambda_n\}$ to the associated Riccati equation. The proof of Theorem 9 depends upon the bound

$$h_n(x) \leq [1 + 2(J_{n-1} - J_n)]h_{n-1}(x) + b_n, \qquad (1.38)$$ e:h-bdds2

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

$$\Lambda_n \leq [1 + 2(J_{n-1} - J_n)]\Lambda_{n-1}, \qquad n \geq 1. \qquad (1.39)$$ e:Pdecrease

It may be shown directly that $\Lambda_n \leq \Lambda_{n-1}$ [19, 60], so the bound (1.38) is not tight in the linear model. However, the semi-decreasing property (1.39) is sufficient to deduce convergence of the algorithm.

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. Theorem 7 shows that it is enough to assume only stability.

1.6 Network models

s:netEx

We now apply the general results of Section 1.4 to the scheduling problem for multiclass queueing 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 [43].

Consider a network composed of d single server stations, indexed by $\sigma = 1, \ldots, d$. 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,

$$x_{t+1} = x_t + \sum_{k=0}^{\ell} I_{t+1}(k)[e^{k+1} - e^k]a_t(k), \tag{1.40}$$

where the state process $\boldsymbol{x}$ evolves on $\mathbb{X} = \mathbb{N}^\ell$, 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 $\{I_n : n \geq 0\}$ are i.i.d. on $\{0,1\}^{\ell+1}$, with

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

For $1 \leq k \leq \ell$, μ_k denotes the service rate for class k customers. For $k = 0$, we let $\mu_0 \stackrel{\Delta}{=} \lambda$ denote the arrival rate of customers of class 1. For $1 \leq k \leq \ell$ we let e^k denote the kth basis vector in $\mathbb{R}^\ell$, and we set $e^0 = e^{\ell+1} \stackrel{\Delta}{=} 0$.

The sequence $\{a_t : t \geq 0\}$ is the control, which takes values in $\{0,1\}^{\ell+1}$. 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 $1 \leq k \leq \ell$, $a_k = 0$ or 1;

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

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

(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 buffer is empty. Condition (iii) means that only one customer may be served at a given instant at a single machine σ. The non-idling condition (iv) is satisfied by any optimal policy with this cost criterion: An inductive proof is given in [41] based upon value iteration.

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 $c(x,a) = |x| \stackrel{\Delta}{=} \sum_k x(k)$. Since $\mathbb{A}(x)$ is a finite set for any x, it follows that (A1) holds with this cost function.

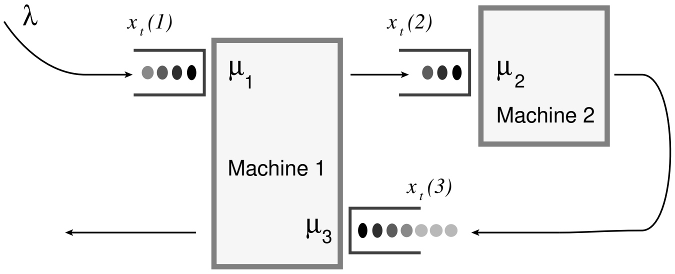


Figure 1.3: A multiclass network with $d = 2$ and $\ell = 3$. f:net

The transition function has the simple form,

$$\begin{aligned}
\mathbb{P}_a(x, x + e^{k+1} - e^k) &= \mu_k a_k, \qquad 0 \le k \le \ell. \\
\mathbb{P}_a(x, x) &= 1 - \sum_0^{\ell} \mu_k a_k
\end{aligned}$$

The accessibility condition (1.4) e:accessible 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.

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

$$\phi^x(t) = \frac{1}{m} x_{mt}.$$

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

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

where the overbar denotes weak closure. This set of stochastic process of course depends on the particular policy π which has been applied. The process ϕ evolves on the state space $\mathbb{R}_+^\ell$ and, for a wide class of scheduling policies, satisfies a differential equation of the form

$$\frac{d}{dt}\phi(t) = \sum_{k=0}^{\ell} \mu_k [e^{k+1} - e^k] u_t(k), \tag{1.41}$$ e:fluidNet

where the function u_t is analogous to the discrete control, and satisfies similar constraints (see the M/M/1 queue model described earlier, or dai95a,daiwei96 [13, 12] for more general examples).

Stability of (e:Net 1.40) in terms of c-regularity is closely connected with the stability of the fluid model dai95a,kummey96a,daimey95a [13, 35, 14]. The fluid model $\mathcal{L}$ is called *L_p-stable* if

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

It is shown in kummey96a [35] that L_2-stability of the fluid model is equivalent to a form of c-regularity for the network.

t:equi **Theorem 11 (Kumar and Meyn kummey96a [35])** *The following stability criteria are equivalent for the network under any nonidling policy.*

(i) The drift condition (e:cFoster 1.9) holds for some function V. The function V is equivalent to a quadratic in the sense that, for some $\gamma > 0$,

$$1 + \gamma|x|^2 \le V(x) \le 1 + \gamma^{-1}|x|^2, \qquad x \in \mathbb{X}. \tag{1.42}$$ e:QuadEquiv

(ii) For some quadratic function V,

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

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

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

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

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

■

The previous result can be strengthened: if the fluid model is L_2-stable, then in fact $\phi(t) = 0$ for all t sufficiently large mey97a [41].

Using this result it is shown in mey97b [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 optimal for both the network and its fluid model.

Let us turn to the VIA: In view of Theorem t:equi 11, how should we initialize the algorithm? Two possibilities are suggested:

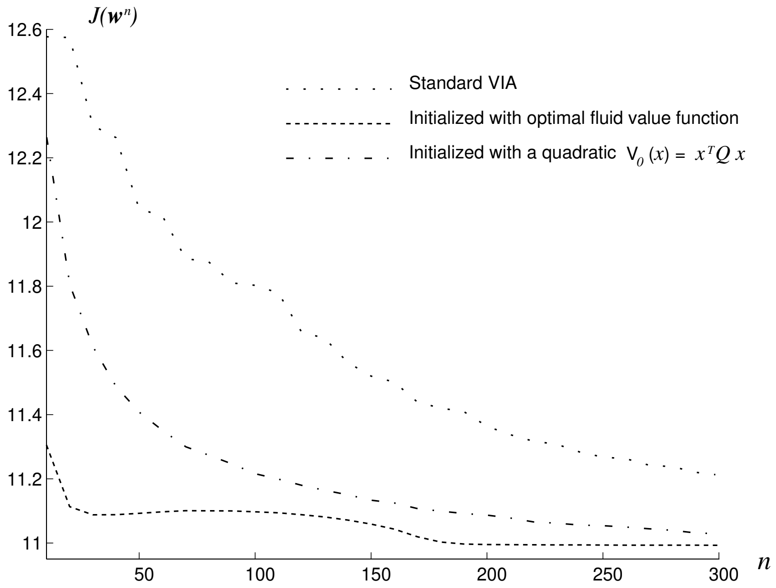


Figure 1.4: Convergence of the VIA with V_0 taken as the value function for the associated fluid control problem, or a pure quadratic function obtained through a linear program. f:Fluid

(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,

$$V_0(x) = \min \int_0^\infty |\phi(t)|\, dt \quad \phi(0) = x, \ x \in \mathbb{X},$$

where the minimum is with respect to all policies for the fluid model. One can show that for large x, V_0 does approximate the relative value function mey97a,mey97b,henmey99a [41, 42, 25].

(ii) The conclusion that the relative value function is 'nearly quadratic' suggests that we search for a pure quadratic form satisfying (e:cFoster 1.9),

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

In kummey96a [35] a linear program is constructed to compute a quadratic solution to (e:cFoster 1.9) for network models, based on prior results of kumkum94a,berpastsi92a [36, 46].

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 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.

Continuous time There is very little of interest to say in the continuous time framework unless one specializes to an interesting class of models. 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

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

with $\beta > 0$. Again one can show that any variable of interest (the invariant measures, solutions to Poisson's equation, or solutions to (1.9)) can be mapped between the resolent and the continuous time process. Further discussion may be found in [42, 53].

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

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

where the 'risk factor' γ is typically assumed to be a small, positive number. Models of this sort were first considered in [3, p. 329] for finite state space models, and complete solutions to the optimization problem were given in [32, 51], again in the finite state space case. This control problem has attracted more recent attention because of the interesting connections between risk sensitive control and game theory [33, 21, 60].

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 very weak, but 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, and there are also simulation based versions of both value and policy iteration (see [4]).

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. In the case of network models, either fluid models or Brownian motion models provide approaches to aggregation which deserve further study.

Bibliography

araborferghomar93 [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.

balmey98a [2] S. Balaji and S.P. Meyn. Multiplicative ergodic theorems for an irreducible Markov chain. *Stoch. Processes and their Appl.*, 1999. Submitted for publication.

bel57 [3] R. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, NJ, 1957.

bertsi96a [4] Bertsekas, D., Tsitsiklis, J. Neuro-Dynamic Programming, Athena Scientific, Belmont, MA, 1996.

bor91 [5] V. S. Borkar. *Topics in controlled Markov chains.* Pitman Research Notes in Mathematics Series # 240, Longman Scientific & Technical, UK, 1991.

bormey98a [6] V. Borkar and S.P. Meyn. Risk Sensitive Optimal Control: Existence and Synthesis for Models with Unbounded Cost. *Mathematics of O.R.*, 1999. Submitted for publication.

cav96 [7] R. Cavazos-Cadena. Value iteration in a class of communicating Markov decision chains with the average cost criterion. Technical report, Universidad Autónoma Agraria Anonio Narro, 1996.

cavfer95c [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.

cavfer98a [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 (1999) 49:299-324.

chemey99a [10] R-R. Chen and S.P. Meyn. Value iteration and optimization of multiclass queueing networks. to appear, QUESTA, 1999.

enghummey96a [11] 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.

daiwei96 [12] 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.

dai95a [13] 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.

daimey95a [14] 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.

dek85a [15] R. Dekker. *Denumerable Markov Decision Chains: Optimal Policies for Small Interest Rates*, PhD thesis, University of Leiden, Leiden, the Netherlands, 1985.

dek87 [16] R. Dekker. Counterexamples for compact action Markov decision chains with average reward criteria. *Comm. Statist.-Stoch. Models*, 3:357–368, 1987.

der66 [17] C. Derman. Dunumerable state MDPs. *Ann. Amth. Statist.*, 37:1545–1554, 1966.

dowmeytwe96a [18] D. Down, S. P. Meyn, and R. L. Tweedie. Geometric and uniform ergodicity of Markov processes. *Ann. Probab.*, 23(4):1671–1691, 1996.

duf90 [19] M. Duflo. *Méthodes Récursives Aléatoires*. Masson, 1990.

dynyus79 [20] 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.

flemce92 [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.

glymey96a [22] P. W. Glynn and S. P. Meyn. A Lyapunov bound for solutions of Poisson's equation. *Ann. Probab.*, 24, April 1996.

hermar96 [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**(1-2), 1998, pp. 105-106.

hen97a [24] Henderson, S. G. *Variance Reduction Via an Approximating Markov Process.* Ph.D. thesis. Department of Operations Research, Stanford University. Stanford, California, USA, 1997.

henmey99a [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.

lerlas96 [26] O. Hernández-Lerma and J. B. Lasserre. Discrete time Markov control processes I. Springer-Verlag, New York, 1996.

hermoncav91 [27] 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.

hor77 [28] A. Hordijk. *Dynamic Programming and Markov Potential Theory.* 1977.

horput87 [29] A. Hordijk and M. L. Puterman. On the convergence of policy iteration. *Math. Op. Res.*, 12:163–176, 1987.

horspitwe95 [30] A. Hordijk, F. M. Spieksma, and R. L. Tweedie. Uniform stability conditions for general space Markov decision processes. Technical report, Leiden University and Colorado State University, 1995. Technical Report.

how60 [31] R. A. Howard. *Dynamic Programming and Markov Processes.* John Wiley and Sons/MIT Press, New York, NY, 1960.

howmat72a [32] R.A. Howard and J.E. Matheson. Risk-sensitive Markov decision processes. *Management Sci.*, 8:356–369, 1972.

jac73 [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.

konbor96 [34] V. R. Konda and V. S. Borkar. Learning algorithms for Markov decision processes. Technical report, Indian Institute of Science, Bangalore, 1996.

kummey96a [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, January 1996.

kumkum94a [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.

kwasiv72 [37] H. Kwakernaak and R. Sivan. *Linear Optimal Control Systems.* Wiley-Interscience, New York, NY, 1972.

lip75 [38] S. Lippman. Applying a new device in the optimization of exponential queueing systems. *Operations Research*, 23:687–710, 1975.

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

meytwe93b [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.

mey97a [41] S.P. Meyn. Stability and optimization of multiclass queueing networks and their fluid models. In *proceedings of the summer seminar on "The Mathematics of Stochastic Manufacturing Systems"*. American Mathematical Society, 1997.

mey97b [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.

mey99a [43] S.P. Meyn. Feedback Regulation for Sequencing and Routing in Multiclass Queueing Networks. submitted for publication, 1999.

num84 [44] E. Nummelin. *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press, Cambridge, 1984.

num91 [45] E. Nummelin. On the Poisson equation in the potential theory of a single kernel. *Math. Scand.*, 68:59–82, 1991.

berpastsi92a [46] I. Ch. Paschalidis D. Bertsimas and J. N. Tsitsiklis. Scheduling of multiclass queueing networks: Bounds on achievable performance. In *Workshop on Hierarchical Control for Real–Time Scheduling of Manufacturing Systems*, Lincoln, New Hampshire, October 16–18, 1992.

per93a [47] 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).

put94 [48] M. L. Puterman. *Markov Decision Processes*. Wiley, New York, 1994.

ritsen93 [49] 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.

ros92 [50] 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.

rot84a [51] U.G. Rothblum. Multiplicative Markov decision chains. *Math. Operations Res.*, 9:6–24, 1984.

sen86 [52] 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.

sen99a [53] L.I. Sennott. Stochastic Dynamic Programming and the Control of Queueing Systems. *Wiley*, 1999.

sen96a [54] L.I. Sennott. The convergence of value iteration in average cost Markov decision chains. *Operations Research Letters*, 19:11–16, 1996.

spi90a [55] F.M. Spieksma. *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*, PhD thesis, University of Leiden, Leiden, the Netherlands, 1990.

tsivan96 [56] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. technical report LIDS-P-2322, Massachusetts Institute of Technology, Cambridge, MA, March 1996. To appear in the IEEE Transactions on Automatic Control.

tuotwe94a [57] P. Tuominen and R.L. Tweedie. Subgeometric rates of convergence of f-ergodic Markov chains. *Adv. Appl. Probab.*, 26:775–798, 1994.

websti87 [58] R. Weber and S. Stidham. Optimal control of service rates in networks of queues. *Adv. Appl. Probab.*, 19:202–218, 1987.

wei94a [59] G. Weiss. On the optimal draining of re-entrant fluid lines. Technical report, Georgia Georgia Institute of Technology and Technion, 1994.

whi90 [60] P. Whittle. *Risk-Sensitive Optimal Control.* John Wiley and Sons, Chichester, NY, 1990.