Risk Sensitive Optimal Control for Markov Decision Processes with Monotone Cost*

V.S. Borkar† S.P. Meyn‡

June 21, 2003

Abstract

The existence of an optimal feedback law is established for the risk sensitive optimal control problem with denumerable state space. The main assumptions imposed are irreducibility, and a *near monotonicity* condition on the one-step cost function. It is found that a solution can be found constructively using either value iteration or policy iteration under suitable conditions on initial feedback law.

Keywords: Optimal Control; Risk Sensitive Control; Dynamic Programming.

1 Introduction

This paper concerns optimal control of Markov Decision Processes (*MDP*s). Formally, this is defined by a triple $(\mathsf{X}, \mathcal{A}, P_a)$ where X is the *state space*, and $\mathcal{A}$ is the *action space*. We assume that both X and $\mathcal{A}$ are denumerable sets. In this case, P_a is, for any $a \in \mathcal{A}$, a transition matrix on the state space X.

*Work supported in part by the Dept. of Science and Technology (Govt. of India) grant no.III5(12)/96-ET., and National Science Foundation grant ECS 940372

†School of Technology and Computer Science, Tata Institute of Fundamental Research. Homi Bhabha Road, Mumbai 400005, INDIA (borkar@tifr.res.in)

‡Coordinated Sciences Laboratory and Department of Electrical and Computer Engg., University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA (s-meyn@uiuc.edu). Part of the research for this paper was done while this author was a Fulbright research scholar and visiting professor at the Indian Institute of Science, and a visiting professor at the Technion.

A sequence $\{u_k\}$ evolving in $\mathcal{A}$ is called an *admissible control sequence* if

$$u_k \in \mathcal{F}_k\,, \qquad k \geq 0,$$

where $\mathcal{F}_k := \sigma\{\Phi_0, \dots, \Phi_k\}$, $k \geq 0$, is the minimal σ-field generated by the *observations*, and the *state process* $\boldsymbol{\Phi}$ is recursively defined via,

$$\mathsf{P}\{\Phi_{k+1} \in A \mid \Phi_0^k; u_0^k; \Phi_k = x; u_k = a\} = P_a(x, A), \qquad x \in \mathsf{X},\ A \subset \mathsf{X},\ a \in \mathcal{A}.$$

We suppose that there is a *one step cost* function $C\colon \mathsf{X} \times \mathcal{A} \to \mathbb{R}_+$, so that for a particular Markov policy $\boldsymbol{w} = (w_0, w_1, w_2, \dots)$ the *risk sensitive cost* starting at $x \in \mathsf{X}$ is defined by

$$R(x, \boldsymbol{w}) := \limsup_{n\to\infty} \frac{1}{n}\Big(\log \mathsf{E}_x^{\boldsymbol{w}}\big[\exp\big(\alpha S_n\big)\big]\Big)\,, \tag{1}$$

where $S_n = \sum_{k=0}^{n-1} C(\Phi_k, w_k(\Phi_k))$, and the expectation above is conditioned on $\Phi_0 = x$. For the processes considered here, the limit supremum in (1) will typically be a limit which is independent of the initial condition $x \in \mathsf{X}$. We consider only the *risk-averse* case where $\alpha > 0$.

Models of this sort were first considered in [2, p. 329] for finite state space models. An in-depth analyis first appeared in [11] in the finite state space case where each controlled chain is irreducible and aperiodic. The general finite state space case was subsequently treated in [20].

There has been renewed interest in the cost criterion (1) during the past decade. The primary reason is the original one: when $\alpha > 0$ the use of the exponential reduces the possibility of rare, but devastating large excursions of the state process. This control problem has attracted more recent attention because of the interesting connections between risk sensitive control and game theory (see [12] or the more recent treatments [15, 22, 7, 6, 8, 13, 23]).

Under certain conditions on the model (in particular, when the model is linear in (x, a)), the controls that optimize (1) are known to be insensitive to specific forms of model uncertainty [22, 6]. In general it may be shown that any stationary policy which gives rise to a finite risk senstive cost will enjoy some attractive properties. The controlled chain is V-uniformly ergodic (see Theorem 3.3), which itself implies some degree of robustness to model uncertainty [9].

The results developed in the present paper are most closely related to [10, 3]. This prior work considers models with bounded cost functions, and imposes a strong form of uniform ergodicity in order to show that a relative value function exists and is bounded. It is also assumed in this

prior work that the constant α appearing in (1) be sufficiently small. The main contribution of this paper is to establish existence of optimal policies under a simple growth condition on the one step cost function. These results hold without any conditions on the 'risk factor' α.

As in [11, 10, 3], we require that each of the controlled chains be irreducible. This can be relaxed to ψ-irreducibility, as defined in [16], with slightly weaker conclusions. However, the general non-irreducible case is subtle, as the treatment [20] of the general finite state space case shows. Fortunately, most models found in applications exhibit some form of irreducibility.

We also show here that stabilizing feedback policies are generated using either the value iteration or the policy iteration algorithm, provided that either algorithm is initialized with a stabilizing feedback law. This generalizes recent results of [17, 4, 18] for the risk neutral ergodic control problem. Under additional assumptions it is shown that either algorithm converges to a solution to the dynamic programming equations.

The remainder of the paper is organized as follows. In the following section we present some background on ergodic theory and the existence of a relative value function for the risk sensitive control problem. Section 3 contains a proof that an optimal policy exists for normlike cost criteria. In Sections 4 and 5 we present analyses of the value iteration and policy iteration algorithms.

2 Multiplicative Ergodic Theorems

In order to address the optimization problem spelled out in the introduction we first state some results from [1] which show that the limit supremum in (1) is in fact a limit when the system is controlled using a stabilizing, stationary policy.

We describe in this section results for a Markov chain without control. We suppose that $\mathbf{\Phi} = \{\Phi_0, \Phi_1, \dots\}$ is an aperiodic and irreducible Markov chain with transition probability P on a countably infinite state space X. We denote by $C\colon \mathsf{X} \to \mathbb{R}_+$ a fixed, non-negative valued function on X, and let $c(x) = \exp(C(x))$, $x \in \mathsf{X}$.

The function C is assumed to be *norm-like*: the sublevel set $\{x : C(x) \le n\}$ is finite for each n [16].

We first present a collection of ergodic theorems from [16]. Theorem 2.1 will be useful below, and it also serves to highlight the symmetry between classical ergodic theory and more recently developed multiplicative ergodic

theory for Markov chains. A Markov chain satisfying the drift criterion (2) with C norm-like and $\mathbf{\Phi}$ irreducible is V-uniformly ergodic (see [16] for notation and related results).

The existence of the two limits in Theorem 2.1 is a consequence of the Geometric Ergodic Theorem of [16]. That the limit $\hat{C}$ is the essentially unique solution to Poisson's equation is discussed on page 433 of [16]. The characterization of the limit γ in (i) is simply the characterization of the steady state mean $\pi(C)$ given in Theorem 10.0.1 of [16].

The first entrance time and first return time to a state θ are defined respectively by

$$\sigma_\theta = \min(k \geq 0 : \Phi_k = \theta); \qquad \tau_\theta = \min(k \geq 1 : \Phi_k = \theta).$$

Theorem 2.1 *Suppose that $\mathbf{\Phi}$ is an irreducible and aperiodic Markov chain with countable state space X and that C is norm-like. Suppose further that there exists $V\colon \mathsf{X} \to [1,\infty)$, and constants $b < \infty$, $\eta < 1$ all satisfying*

$$\mathsf{E}_x[V(\Phi_1)] = \sum_{y\in\mathsf{X}} P(x,y)V(y) \leq \eta V(x) - C(x) + b. \tag{2}$$

Then there exists a constant $\gamma \in \mathbb{R}_+$ and a function $\hat{C}\colon \mathsf{X} \to \mathbb{R}$ such that

$$\lim_{n\to\infty} \mathsf{E}_x\big[S_n - \gamma n\big] = \hat{C}(x), \quad \textit{and hence,} \ \lim_{n\to\infty} \frac{1}{n}\mathsf{E}_x\big[S_n\big] = \gamma,$$

where

(i) *The constant γ is the unique solution to*

$$\mathsf{E}_\theta\Big[\sum_{k=0}^{\tau_\theta - 1}\Big(C(\Phi_k) - \gamma\Big)\Big] = 0.$$

(ii) *The function $\hat{C}$ solves the Poisson equation*

$$P\hat{C}\,(x) = \hat{C}(x) - C(x) + \gamma, \qquad x \in \mathsf{X}.$$

(iii) *The solution $\hat{C}$ is unique up to an additive constant: If $\hat{C}'$ is any other solution, then*

$$\hat{C}(x) - \hat{C}(x_0) = \hat{C}'(x) - \hat{C}'(x_0), \qquad x, x_0 \in \mathsf{X}.$$

□

The desired multiplicative ergodic theorem is expressed in the following result, which is evidently closely related to Theorem 2.1. This and some related results are developed in [1].

Theorem 2.2 *Suppose that $\mathbf{\Phi}$ is an irreducible and aperiodic Markov chain with countable state space X, and that C is norm-like. Suppose further that there exists $V_0 \colon \mathsf{X} \to \mathbb{R}_+$, and constants $B < \infty$, $\alpha_0 > 0$ all satisfying*

$$\mathsf{E}_x[\exp(V_0(\Phi_1))] = \sum_{y \in \mathsf{X}} P(x,y) \exp\bigl(V_0(y)\bigr) \le \exp\bigl(V_0(x) - \alpha_0 C(x) + B\bigr). \quad (3)$$

Then there exists a (possibly infinite) constant $\overline{\alpha} \ge \alpha_0$, and a convex, increasing function $\Lambda : \mathbb{R} \to \mathbb{R}$ such that $\Lambda(\alpha) < \infty$ for $\alpha < \overline{\alpha}$; and $\Lambda(\alpha) = \infty$ for $\alpha > \overline{\alpha}$. Furthermore, the following hold:

For any $\alpha < \overline{\alpha}$, there is a function $\check{c}_\alpha \colon \mathsf{X} \to \mathbb{R}_+$ such that

$$\lim_{n \to \infty} \mathsf{E}_x\bigl[\exp\bigl(\alpha S_n - n\Lambda(\alpha)\bigr)\bigr] = \check{c}_\alpha(x), \quad (4)$$

and for all *α,*

$$\lim_{n \to \infty} \frac{1}{n} \log\Bigl(\mathsf{E}_x\bigl[\exp\bigl(\alpha S_n\bigr)\bigr]\Bigr) = \Lambda(\alpha).$$

Moreover, for all $\alpha < \overline{\alpha}$,

(i) *the constant $\Lambda(\alpha) \in \mathbb{R}$ is the unique solution to*

$$\mathsf{E}_\theta\Bigl[\exp\Bigl(\sum_{k=0}^{\tau_\theta - 1} \alpha C(\Phi_k) - \Lambda(\alpha)\Bigr)\Bigr] = 1.$$

(ii) *The function $\check{c}_\alpha$ solves the multiplicative Poisson equation:*

$$P\check{c}_\alpha\,(x) = \check{c}_\alpha(x) \exp\bigl(-\alpha C(x) + \Lambda(\alpha)\bigr), \qquad x \in \mathsf{X}. \quad (5)$$

(iii) *The solution $\check{c}_\alpha$ is unique up to constant multiples: If $\check{c}'_\alpha$ is any other solution, then*

$$\frac{\check{c}_\alpha(x)}{\check{c}_\alpha(x_0)} = \frac{\check{c}'_\alpha(x)}{\check{c}'_\alpha(x_0)}, \qquad x, x_0 \in \mathsf{X}.$$

□

Analogous results for a bounded function C are also obtained in [1].

The constant $\lambda(\alpha) = \exp(\Lambda(\alpha))$ is equal to the *generalized principal eigenvalue* (g.p.e.) for the kernel $\widehat{P}_\alpha$ defined by

$$\widehat{P}_\alpha(x,y) = \exp(\alpha C(x)) P(x,y), \qquad x, y \in \mathsf{X}.$$

It is also known as the *Perron-Frobenius eigenvalue*, and $R(\alpha) = \lambda(\alpha)^{-1}$ is the convergence parameter (see [1]).

The function $\check{c}_\alpha$ is the corresponding Perron-Frobenius eigenfunction for $\widehat{P}$, and (5) is a restatement of the eigenfunction equation $\widehat{P}_\alpha \check{c} = \lambda(\alpha)\check{c}$. The term *multiplicative Poisson equation* is used to stress the symmetry with the previous theorem, and with the usual MDP theory under the average cost optimality criterion.

Suppose that the Markov chain is recurrent, as it will be under (2) or (3). For any α, the constant $\Lambda(\alpha) = \log(\lambda(\alpha))$ is given by the following formula:

$$\Lambda(\alpha) := \inf\big\{\Lambda \in \mathbb{R} : \mathsf{E}_\theta\big[\exp\big(\alpha S_{\tau_\theta} - \tau_\theta \Lambda\big)\big] \le 1\big\}, \tag{6}$$

with θ equal to any fixed state in X.

From the definition of $\Lambda(\alpha)$ and Fatou's Lemma we have, whenever $\Lambda(\alpha) < \infty$,

$$\xi(\alpha) := \mathsf{E}_\theta\big[\exp\big(\alpha S_{\tau_\theta} - \tau_\theta \Lambda(\alpha)\big)\big] \le 1. \tag{7}$$

The constant $\overline{\alpha}$ is then defined as $\overline{\alpha} = \sup\{\alpha : \xi(\alpha) = 1\}$ [1]. It is shown there that for any $\alpha < \overline{\alpha}$, the function

$$h^\alpha(x) = \mathsf{E}_x\big[\exp\big(\alpha S_{\tau_\theta} - \tau_\theta \Lambda(\alpha)\big)\big], \qquad x \in \mathsf{X},$$

is the unique (up to constant multiples) solution to the multiplicative Poisson equation. The function h^α will appear as the relative value function for the optimization problems considered below.

The drift criterion (3) is useful since it gives a bound on $\overline{\alpha}$, and it also implies a strong form of ergodicity for the chain. It is equivalent to the following 'sub-eigenvector equation',

$$\widehat{P}_{\alpha_0} V(x) := \exp(\alpha_0 C(x)) \sum_{y \in \mathsf{X}} P(x,y) V(y) \le \lambda V(x), \qquad x \in \mathsf{X}, \tag{8}$$

where $V = \exp(V_0)$, and $\lambda = \exp(B)$.

Using these ideas we find that a solution V to (8) or (3) always exists provided that the "cost" $\Lambda(\alpha)$ is finite. For a proof see [1].

Lemma 2.3 *For an irreducible Markov chain $\mathbf{\Phi}$ and a norm-like function C, the following are equivalent for any $0 < \lambda < \infty$, and any $\alpha > 0$,*

(a) $\mathbf{\Phi}$ *is recurrent and the g.p.e. satisfies*

$$\lambda(\alpha) \le \lambda.$$

(b) *There exists a function $V\colon \mathsf{X} \to \mathbb{R}_+$ satisfying (8), and in addition*

$$\inf_{x\in\mathsf{X}} V(x) > 0. \tag{9}$$

□

The proof of Theorem 2.2 involves a change of measure performed using a solution of the multiplicative Poisson equation (5). We sketch the main ideas here since this change of measure will also be required in some of the results below. For $\alpha < \overline{\alpha}$ define

$$\check{P}_\alpha(x,y) = \frac{\exp(\alpha C(x) - \Lambda(\alpha))}{h^\alpha(x)} P(x,y) h^\alpha(y).$$

where h^α is any solution to the multiplicative Poisson equation which is not identically zero. The kernel $\check{P}_\alpha$ is probabilistic ($\check{P}_\alpha(x,\mathsf{X}) = 1$ for $x \in \mathsf{X}$) since the multiplicative Poisson equation holds. It follows that $\check{P}_\alpha$ is the transition kernel for some Markov chain $\check{\mathbf{\Phi}}^\alpha$. Theorem 2.4 establishes ergodicity of these Markov chains, and shows that $\mathbf{\Phi}$ itself is V-uniformly ergodic when $\overline{\alpha} > 0$.

Theorem 2.4 *For any $\alpha < \overline{\alpha}$ the Markov chain $\check{\mathbf{\Phi}}^\alpha$ is V_α-uniformly ergodic for some $V_\alpha \ge 1$. Hence there is an invariant probability measure $\check{\pi}_\alpha$ for $\check{P}_\alpha$, and for any $g\colon \mathsf{X} \to \mathbb{R}$ satisfying*

$$\Big|\frac{g(x)}{h^\alpha(x)}\Big| \le V_\alpha(x), \qquad x \in \mathsf{X},$$

the following limit holds at a geometric rate as $n \to \infty$:

$$\mathsf{E}_x\big[\exp\big(\alpha S_n - n\Lambda(\alpha)\big) g(\Phi_n)\big] \to h^\alpha(x)\check{\pi}_\alpha(g/h^\alpha), \qquad x \in \mathsf{X}.$$

Proof It is shown in [1] that the chain $\check{\mathbf{\Phi}}^\alpha$ is V_α-uniformly ergodic for some V_α provided that $\alpha < \bar{\alpha} := \sup(\alpha : \Lambda(\alpha) < \infty)$.

The limit then follows from ergodicity and the formula

$$\check{\mathsf{E}}^\alpha_x[f(\check{\Phi}^\alpha_n)] = \frac{1}{h^\alpha(x)} \mathsf{E}_x\big[\exp\big(\alpha S_n - n\Lambda(\alpha)\big) h^\alpha(\Phi_n) f(\Phi_n)\big],$$

valid for any integrable $f\colon \mathsf{X} \to \mathbb{R}$ (see the Geometric Ergodic Theorem of [16]). □

There are several possible extensions of these results: the conditions on the 'cost function' C can be generalized in various directions. One direction which is developed in [17] for the risk neutral control problem is to assume

that the sublevel sets of $C(\cdot)$ are *petite*, as defined in [16], rather than finite or compact. Such conditions may be used to generalize results of the form developed here to arbitrary state spaces [14].

Some extensions are possible even in the countable state space setting. To remove the unboundedness condition on C one may replace the norm-like assumption with *near-monotonicity*, so that $\{x : C(x) \le \eta\}$ is a finite set for any $\eta < \sup_{x\in\mathsf{X}} C(x)$. A parallel ergodic theory is developed in [1] for near-monotone cost functions, and using these results it is possible to generalize all of the results in this paper. For the sake of brevity we do not consider in detail such extensions.

3 Existence of Optimal Controls

We may now address the question of existence of optimal controls for a controlled Markov chain with transition function P_a using the cost criterion (1). We assume without loss of generality that $\alpha = 1$, so that the goal is to minimize over all controls,

$$R(x, \boldsymbol{w}) = \limsup_{n\to\infty} \frac{1}{n} \log\Big(\mathsf{E}_x^{\boldsymbol{w}} \big[\exp\big(S_n^{\boldsymbol{w}} \big) \big] \Big) ,$$

where we set

$$S_n^{\boldsymbol{w}} := \sum_{k=0}^{n-1} C(\Phi_k, w_k(\Phi_k)).$$

The function C is the one-step cost, which is assumed to satisfy a norm-like condition. We let $c(x, a) = \exp(C(x, a))$, and for any function $w\colon \mathsf{X} \to \mathcal{A}$ we write,

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

The function w is interpreted as a *feedback law* in the results below, and the control $u_k = w(\Phi_k)$ is called a *stationary Markov policy.* The control sequence is called *Markov* if $u_k = w_k(\Phi_k)$, $k \ge 0$, for a sequence of functions $\boldsymbol{w} = \{w_k\}$.

Throughout the remainder of this paper we also impose the following assumptions on the state space, action space, and on the controlled chain.

(A1) the state space X is countably infinite; the action space $\mathcal{A}$ is finite; and the function $C(\,\cdot\,, a)$ is norm-like for any fixed $a \in \mathcal{A}$.

(A2) For any Markov policy $\boldsymbol{w}$

$$\mathsf{P}^{\boldsymbol{w}}\{\tau_y < \infty \mid \Phi_0 = x\} > 0, \qquad x, y \in \mathsf{X}.$$

For any stationary policy $\boldsymbol{w}$, the Markov chain with law $\mathsf{P}^{\boldsymbol{w}}$ is assumed to be aperiodic.

It will be clear that the strong assumption on $\mathcal{A}$ used in (A1) can be replaced by appropriate continuity conditions. The norm-like condition on the cost function is more difficult to remove, but some extensions were described in the previous section. Condition (A2) is just an extension of the usual definition of irreducibility for a time homogeneous Markov chain on X.

We now give a generalization of the g.p.e. defined by (6). Let θ be some arbitrary state in X, and for any Markov policy $\boldsymbol{w} = (w_0, w_1, w_2, \dots)$ let

$$\Lambda(\boldsymbol{w}) := \inf\Big\{\Lambda \in \mathbb{R} : \mathsf{E}_\theta^{\boldsymbol{w}}\Big[\exp\Big(\sum_{k=0}^{\tau_\theta - 1}[C(\Phi_k, w_k(\Phi_k)) - \Lambda]\Big)\Big] \le 1\Big\}.$$

The minimal value is denoted

$$\Lambda^* := \inf \Lambda(\boldsymbol{w}), \tag{10}$$

where the infimum is over all Markov policies. For any policy, we let $\lambda(\boldsymbol{w}) = \exp(\Lambda(\boldsymbol{w}))$, and we set $\lambda^* = \exp(\Lambda^*)$.

If $\boldsymbol{w}$ is stationary then we set $\lambda(w) = \lambda(\boldsymbol{w})$, $\Lambda(w) = \log(\lambda(w))$, and in this case, the constant $\lambda(w)$ is the g.p.e. for the kernel

$$\widehat{P}_w(x, y) = c_w(x) P_w(x, y). \tag{11}$$

We call the controlled Markov chain $\boldsymbol{\Phi}^{\boldsymbol{w}}$ *stable* if $\Lambda(\boldsymbol{w}) < \infty$. If $\boldsymbol{w} = (w, w, \dots)$ is stationary then the feedback law w is called *stabilizing*.

Proposition 3.1 shows that $\Lambda(w)$ is indeed the steady state cost when w is a stabilizing feedback law. This is an immediate consequence of Theorem 2.2.

Proposition 3.1 *Suppose that (A1) holds and that $\boldsymbol{w} = (w, w, w, \dots)$ is a stationary policy defined through the stabilizing feedback law w. Then for every initial condition x,*

$$R(x, \boldsymbol{w}) = \lim_{n\to\infty} \frac{1}{n} \log\Big(\mathsf{E}_x\big[\exp\big(S_n^{\boldsymbol{w}}\big)\big]\Big) = \Lambda(w).$$

□

The following bounds are taken from [18].

Lemma 3.2 *Under the assumptions of this section,*

(i) *There exists a function* $s\colon \mathsf{X}\times\mathsf{X} \to (0,1)$ *such that for any Markov policy* $\boldsymbol{w} = (w_0, w_1, \dots)$,

$$K_{\boldsymbol{w}}(x,y) := \mathsf{E}_x^{\boldsymbol{w}}\Big[\sum_{k=0}^{\infty} 2^{-(k+1)}\mathbb{I}(\Phi_k = y)\Big] \geq s(x,y), \qquad x,y \in \mathsf{X}.$$

(ii) *For any finite set* $S \subset \mathsf{X}$ *and any* $y \in \mathsf{X}$, *there is a finite constant* $B = B(S,y)$ *such that for any Markov policy,*

$$\mathsf{E}_x^{\boldsymbol{w}}\Big[\sum_{k=0}^{\tau_y - 1} \mathbb{I}_S(\Phi_k)\Big] \leq B, \qquad x \in \mathsf{X}.$$

□

The following result illustrates that stability implies a strong form of ergodicity for the chain.

Theorem 3.3 *If* w *is a stabilizing feedback law then the controlled chain* $\boldsymbol{\Phi}^w$ *is* V*-uniformly ergodic for some* V *satisfying* $c_w(x) \leq V(x)$, $x \in \mathsf{X}$.

Hence, in particular,

(i) *the chain is ergodic with unique invariant probability* π_w.

(ii) *There exists* $\rho < 1$, $B_0 < \infty$, *such that for any function* f *satisfying* $|f| \leq V$,

$$|\mathsf{E}_x^w[f(\Phi_n)] - \pi_w(f)| \leq B_0 V(x)\rho^n, \qquad x \in \mathsf{X}, n \geq 0.$$

Proof If the feedback law is stabilizing then we have seen in Lemma 2.3 that there is a function $V \geq 1$ such that

$$c_w P_w V \leq \lambda(w) V.$$

It then follows that $V \geq \lambda(w)^{-1} c_w$, so that the bound $V \geq c_w$ can be obtained on scaling V. Letting S denote the finite set $S = \{x : \lambda(w)^{-1} c_w(x) \leq 2\}$ we obtain, for some $b < \infty$,

$$P_w V \leq (1/2)V + b\mathbb{I}_S,$$

which establishes V-geometric ergodicity (see Theorem 16.0.1 of [16]). □

For general Markov policies we cannot exactly duplicate Proposition 3.1 but we can obtain a lower bound.

Proposition 3.4 *Under (A1) and (A2),* $R(x, \boldsymbol{w}) \geq \Lambda^*$ *for any Markov policy* $\boldsymbol{w}$, *and any* $x \in \mathsf{X}$.

Proof Let $0 < \Lambda < \Lambda^*$ be arbitrary. For any Markov policy $\boldsymbol{w}$ we must then have

$$\mathsf{E}_\theta^{\boldsymbol{w}}\Big[\exp\Big(\sum_{k=1}^{\tau_\theta}[C(\Phi_k, w_k(\Phi_k)) - \Lambda]\Big)\Big] > 1.$$

Using Fatou's Lemma we may assert the existence of $N_0 \geq 1$ such that for any Markov policy and any $N \geq N_0$,

$$\mathsf{E}_\theta^{\boldsymbol{w}}\Big[\exp\Big(\sum_{k=1}^{\tau_\theta \wedge N}[C(\Phi_k, w_k(\Phi_k)) - \Lambda]\Big)\Big] \geq 1. \tag{12}$$

For $N \geq N_0$ let

$$W_N(x) = \min \mathsf{E}_x^{\boldsymbol{w}}\Big[\exp\Big(\sum_{k=0}^{\sigma_\theta \wedge (N-1)}[C(\Phi_k, w_k(\Phi_k)) - \Lambda]\Big)\Big], \tag{13}$$

where the minimum is taken over all Markov policies.

Fix N, and suppose that the minimum is achieved at $\bar{\boldsymbol{w}}$. Using Jensen's inequality we have

$$\log(W_N(x)) \geq -\Lambda \mathsf{E}_x^{\bar{\boldsymbol{w}}}\Big[\sum_{k=0}^{\sigma_\theta}\mathbb{I}_S(\Phi_k)\Big]$$

where $S = \{x : \min_a C(x,a) \leq \Lambda\}$ is finite. By (13) and Lemma 3.2 (ii) we see that $W_N(x) \geq \delta := \exp(-\Lambda B(S,\theta))$ for all x and N.

For any feedback law w we have

$$\lambda^{-1}c_w(x)P_w W_N(x) = \mathsf{E}_x^{\bar{\boldsymbol{w}}'}\Big[\exp\Big(\sum_{k=0}^{\tau_\theta \wedge N}[C(\Phi_k, \overline{w}_{k-1}(\Phi_k)) - \Lambda]\Big)\Big],$$

where $\bar{\boldsymbol{w}} = (\overline{w}_0, \overline{w}_1, \overline{w}_2, \dots)$, $\bar{\boldsymbol{w}}' = (w, \overline{w}_0, \overline{w}_1, \dots)$. From the definition of $(W_N : N \geq N_0)$ and (12) we then have

$$\lambda^{-1}c_w(x)P_w W_N(x) \geq W_{N+1}(x)$$

We note that the bound (12) covers the case where $x = \theta$. Since the feedback law w is arbitrary we may iterate the previous bound to obtain for any Markov policy $\boldsymbol{w}$,

$$\lambda^{-n}\mathsf{E}_x^{\boldsymbol{w}}\Big[\exp\Big(\sum_{k=0}^{n-1}C(\Phi_k, w_k(\Phi_k))\Big)W_{N_0}(\Phi_n)\Big] \geq W_{N_0+n}(x) \geq \delta.$$

From the Markov property and minimality of W_{N_0} we then obtain the bound

$$\lambda^{-n}\mathsf{E}_x^{\boldsymbol{w}}\Big[\exp\Big(\sum_{k=0}^{n+N_0-1} C(\Phi_k, w_k(\Phi_k))\Big)\Big] \geq \delta.$$

In conclusion, we see that

$$\liminf_{n\to\infty} \mathsf{E}_x^{\boldsymbol{w}}\big[\exp\big(\sum_{k=0}^{n-1}[C(\Phi_k, w_k(\Phi_k) - \Lambda])\big)\big] \geq \lambda^{-N_0}\delta > 0.$$

Hence $R(x, \boldsymbol{w}) \geq \Lambda^*$ since $\Lambda < \Lambda^*$ is arbitrary. □

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

$$h_*(x) \;:=\; \inf_{\boldsymbol{w}} \mathsf{E}_x^{\boldsymbol{w}}\Big[\exp\Big(\sum_{k=0}^{\sigma_\theta}[C(\Phi_k, w_k(\Phi_k)) - \Lambda^*]\Big)\Big] \tag{14}$$

$$w^*(x) \;:=\; \arg\min_{a\in\mathcal{A}} c(x,a)P_a h_*\,(x), \tag{15}$$

where in (15) the policy w^* is taken to be any solution to the minimization.

Lemma 3.5 *If Λ^* is finite, then*

(i) *The function h_* is everywhere finite.*

(ii) *The multiplicative Poisson inequality holds,*

$$c_{w^*}(x)P_{w^*}h_*\,(x) \leq \lambda^* h_*(x), \qquad x \in \mathsf{X}. \tag{16}$$

Proof We first show that there exists a Markov policy $\underline{\boldsymbol{w}}$ such that $\Lambda(\underline{\boldsymbol{w}}) = \Lambda^*$.

Take a sequence $\{\Lambda_n\}$ for which $\Lambda_n \downarrow \Lambda^*$ as $n \downarrow \infty$, and choose Markov policies $\boldsymbol{w}^n$ for which $\Lambda(\boldsymbol{w}^n) \leq \Lambda_n$ for each n. We then have,

$$\mathsf{E}_\theta^{\boldsymbol{w}^n}\Big[\exp\Big(\sum_{k=0}^{\tau_\theta-1}[C(\Phi_k, w_k^n(\Phi_k)) - \Lambda_n]\Big)\Big] \leq 1.$$

Assume that there is a Markov policy $\boldsymbol{w}^\infty$ such that $\boldsymbol{w}^n \to \boldsymbol{w}^\infty$ as $n \to \infty$ pointwise. This is possible on taking a subsequence since the control set is finite.

Pointwise convergence is equivalent to weak convergence on $\mathsf{X}^{\mathbb{Z}_+}$. Since c is positive we then obtain,

$$\mathsf{E}_\theta^{\boldsymbol{w}^\infty}\Big[\exp\Big(\sum_{k=0}^{\tau_\theta-1}[C(\Phi_k, w_k^\infty(\Phi_k)) - \Lambda^*]\Big)\Big] \le 1,$$

so that $\Lambda(\boldsymbol{w}^\infty) \le \Lambda^*$. By minimality of Λ^* this must be an equality, and hence we may take $\underline{\boldsymbol{w}} = \boldsymbol{w}^\infty$.

Observe that for each $x \ne \theta$,

$$\begin{aligned}
&(\lambda^*)^{-1} c_{w^*}(x) P_{w^*} h_*(x) \\
&\quad = (\lambda^*)^{-1} \min_a \sum_{y\in\mathsf{X}} c(x,a) P_a(x,y) \Big\{ \inf_{\boldsymbol{w}} \mathsf{E}_y^{\boldsymbol{w}} \Big[\exp\big(S_{\sigma_\theta}^{\boldsymbol{w}} - \sigma_\theta \Lambda^*\big)\Big]\Big\} \qquad (17)\\
&\quad = h_*(x)\,,
\end{aligned}$$

while for $x = \theta$ we have

$$P_{w^*} h_*(\theta) = \min_{\boldsymbol{w}} \mathsf{E}_\theta^{\boldsymbol{w}} \Big[\exp\big(S_{\tau_\theta}^{\boldsymbol{w}} - \tau_\theta \Lambda^*\big)\Big] \le 1.$$

It follows that the sub-eigenvector equation (16) holds, which establishes (ii).

One may infer from the inequality (16) that the set $S = \{x \in \mathsf{X} : h_*(x) < \infty\}$ is absorbing. That is, $P_{w^*}(x, S) = 1$ for $x \in S$. Since the point θ is in S, and since the kernel P_{w^*} is irreducible, it follows that $S = \mathsf{X}$, which establishes (i). □

Theorem 3.6 *Suppose that (A1) and (A2) hold, and that $\Lambda^* < \infty$. Then*

(i) *The feedback law w^* is stabilizing with g.p.e. λ^*;*

(ii) *The stationary policy $\boldsymbol{w}^* = (w^*, w^*, w^*, ...)$ is optimal over all Markov policies: For any Markov policy $\boldsymbol{w}$,*

$$R(x, \boldsymbol{w}) \ge R(x, \boldsymbol{w}^*) = \Lambda^*, \qquad x \in \mathsf{X};$$

(iii) *The relative value function h_* is uniformly bounded from below:*

$$\inf_{x\in\mathsf{X}} h_*(x) > 0.$$

Proof Result (iii) follows from (16) which may be used to establish the bound

$$\mathsf{E}_x^{w^*}\Big[\exp\Big(\sum_{k=0}^{\tau_\theta-1} C_{w^*}(\Phi_k) - \Lambda^*\Big)\Big] \le h_*(x)/h_*(\theta), \qquad x \in \mathsf{X}.$$

We then obtain, exactly as in the derivation of the lower bound on W_n defined in (13),

$$h_*(x) \ge h_*(\theta)\exp(-\Lambda^* B(S,\theta)), \qquad x \in \mathsf{X},$$

with $S = \{x \in \mathsf{X} : \min_a C(x,a) \le \Lambda^*\}$.

That w^* is stabilizing with g.p.e. λ^* then follows from (16), Proposition 3.1 and Lemma 2.3, giving (i), and then (ii) follows from Proposition 3.1 and Proposition 3.4. □

Note that the theorem *does not* say that the pair (λ^*, h_*) solves the dynamic programming equations

$$c(x,a)P_a h_*(x) \ge \min_{a\in\mathcal{A}}\big\{c(x,a)P_a h_*(x)\big\} = \lambda^* h_*(x), \qquad x \in \mathsf{X}, a \in \mathcal{A}. \quad (18)$$

The difficulty is that we do not know in general if (16) is in fact an equality. It *is* an equality for all $x \ne \theta$, but for $x = \theta$ the equality can fail (see (7)). This corresponds to the 'R-transient' case for the kernel $c_{w^*}P_{w^*}$ [21]. Fortunately, the inequality provides an upper bound which is enough to show that w^* is optimal.

4 Value Iteration

In this and the following section we assume that the conditions of Theorem 3.6 are met so that an optimal policy exists. We now focus on computational approaches for constructing an optimal stationary feedback law w^*. We first consider the value iteration algorithm, or VIA.

The VIA for the risk sensitive optimal control problem recursively constructs a sequence of value functions $\{V_n : n \ge 0\}$ as follows: For $n = 0$ the function $V_0 \colon \mathsf{X} \to [1,\infty)$ is given as an initial condition. For $n \ge 1$ the value function is defined recursively,

$$V_n(x) = \min_{a\in\mathcal{A}}\big\{c(x,a)P_a V_{n-1}(x)\big\}, \qquad x \in \mathsf{X}.$$

We follow [4] and assume that V_0 is a "Lyapunov function" in the sense of (8) for at least one policy so that for some $\overline{\lambda}_{-1} < \infty$ and one feedback law w_{-1},

$$c_{w_{-1}}(x)P_{w_{-1}}V_0(x) \le \overline{\lambda}_{-1}V_0(x), \qquad x \in \mathsf{X}. \quad (19)$$

For each n we fix a feedback law w^n which achieves the minimum,

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

From the sequence $\{w^n : n \ge 0\}$ of feedback laws we define two policies:

$$\boldsymbol{w}^n = (w^n, w^n, w^n, \dots) \qquad \boldsymbol{v}^n = (w^{n-1}, w^{n-2}, \dots, w^1, w^0, w^0, w^0, \dots).$$

We will find that the feedback law w^n is stabilizing for any n, and that it is near optimal when n is large. The Markov policy $\boldsymbol{v}^n$ minimizes the finite-horizon cost criterion,

$$\mathsf{E}_x^{\boldsymbol{w}} \big[\exp\big(S_n^{\boldsymbol{w}}\big) V_0(\Phi_n) \big],$$

over all Markov policies $\boldsymbol{w}$.

The normalized value function and the incremental cost are defined respectively as

$$h_n(x) = V_n(x)/V_n(\theta); \qquad g_n(x) = V_{n+1}(x)/V_n(x), \qquad x \in \mathsf{X}, n \ge 0.$$

For each n we let $\overline{\lambda}_n := \sup_{x\in\mathsf{X}} g_n(x)$, and $\overline{\Lambda}_n = \log(\overline{\lambda}_n)$. We let $P_n = P_{w^n}$ and $c_n = c_{w^n}$.

Lemma 4.1 *Suppose that (A1) holds and that the initial condition V_0 is chosen so that (19) holds. Then,*

(i) *For each n, the function V_n is bounded from below by unity, and the following inequality holds:*

$$c_n P_n V_n \le \overline{\lambda}_n V_n$$

(ii) *The upper bounds $\{\overline{\lambda}_n\}$ are finite and decreasing:*

$$\overline{\lambda}_{-1} \ge \overline{\lambda}_0 \ge \overline{\lambda}_1 \ge \cdots$$

Proof By definition of w^n we have

$$V_{n+1} = c_n P_n V_n.$$

Hence if $V_n(x) \ge 1$ for all x then $V_{n+1}(x) \ge (\inf_x c_n(x))(\inf_x V_n(x)) \ge 1$ for all x. Since the initial condition V_0 is assumed to be bounded from below

by unity, we see by induction that each V_n is similarly bounded. The proof of (i) is concluded on noting that

$$c_n P_n V_n = g_n V_n \le \overline{\lambda}_n V_n.$$

To prove (ii) observe that for any n,

$$\begin{aligned} g_n = \frac{V_{n+1}}{V_n} &= \frac{c_n P_n V_n}{V_n} \\ &\le \frac{c_{n-1} P_{n-1} V_n}{V_n} \\ &= \frac{c_{n-1} P_{n-1} (g_{n-1} V_{n-1})}{V_n} \\ &\le \overline{\lambda}_{n-1} \frac{c_{n-1} P_{n-1} V_{n-1}}{V_n} = \overline{\lambda}_{n-1} \end{aligned}$$

This shows that the sequence $\{\overline{\lambda}_n\}$ is decreasing. □

From the lemma we find that the value iteration algorithm generates stabilizing policies, provided that it is properly initialized. This is summarized in the following theorem:

Theorem 4.2 *Suppose that (A1) and (A2) hold, and that (19) also holds for some initial feedback law w_{-1} and a finite constant λ_{-1}. Then each of the feedback laws $\{w_n\}$ is stabilizing, and the risk sensitive cost satisfies*

$$R(x, \boldsymbol{w}^n) = \Lambda_n \le \overline{\Lambda}_n < \infty, \qquad x \in \mathsf{X}, n \ge 0.$$

Proof From Lemma 4.1 we have the bound

$$P_n V_n \le \overline{\lambda}_n c_n^{-1} V_n.$$

Hence the result follows from Lemma 2.3 and Proposition 3.1. □

Define inductively a new sequence of functions $\{\widetilde{h}_n\}$ as follows: For $n = 0$ we take $\widetilde{h}_0 = h_0$, and for $n \ge 1$ define

$$\widetilde{h}_n := \frac{1}{\lambda^*} \min_a \big\{ c(x,a) P_a \widetilde{h}_{n-1}(x) \big\}, \qquad x \in \mathsf{X}. \tag{20}$$

By induction we see that $\widetilde{h}_n$ and h_n are constant multiples for each n, and we have the following interpretation:

$$\widetilde{h}_n(x) = (\lambda^*)^{-n} \min_{\boldsymbol{w}} \mathsf{E}_x^{\boldsymbol{w}} \big[\exp\big(S_n^{\boldsymbol{w}}\big) h_0(\Phi_n) \big], \tag{21}$$

where the minimum is over all Markov policies.

To obtain an upper bound on $\{\widetilde{h}_n\}$ we use the following assumptions.

(A3) There exists a solution (λ^*, h_*) to the dynamic programming equations (18) satisfying $h_*(\theta) = 1$, with λ^* given in (10).

(A4) There exists a solution w^* to (15) such that the transformed kernel

$$\check{P}_*(x,y) = \frac{c_{w^*}(x) P_{w^*}(x,y) h_*(y)}{\lambda^* h_*(x)}$$

is positive recurrent with unique invariant probability $\check{\pi}_*$.

We denote the transition kernel P_{w^*} by P_*.

Suppose that w^* is the feedback law defined in (15), with h_* given in (14). If the the Markov chain with transition function P_* satisfies $\Lambda(w^*, \alpha) < \infty$ for some $\alpha > 1$, it then follows from Theorem 2.2 that (A4) holds with $\check{P}_*$ geometrically recurrent (see also [1]). Assumption (A3) will also hold since the multiplicative Poisson equation $c_{w^*} P_* h_* = \lambda^* h_*$ is solved uniquely, again by Theorem 2.2, and by definition of w^* we do have

$$\min_a c(x,a) P_a h_* = c_*(x) P_* h_*(x) = \lambda^* h_*(x), \qquad x \in \mathsf{X}.$$

Lemma 4.3 *Under (A1)–(A4), provided that $\check{\pi}_*(h_0/h_*) < \infty$, the following bounds hold for all initial x:*

$$\begin{aligned} \limsup_{n\to\infty} \widetilde{h}_n(x) &\le \check{\pi}_*(h_0/h_*) h_*(x). \\ \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \log(g_n(x)) &= \Lambda^* \end{aligned}$$

Proof Substituting $\boldsymbol{w}^*$ for $\boldsymbol{w}$ in (21) gives the upper bound,

$$\widetilde{h}_n(x) \le \mathsf{E}_x^{w^*}\Big[\exp\Big(\sum_{k=0}^{n-1} [C(\Phi_k, w^*(\Phi_k) - \Lambda^*]\Big) h_0(\Phi_n)\Big] = h_*(x) \check{\mathsf{E}}_x^*\Big[\frac{h_0}{h_*}(\Phi_n)\Big].$$

From the f-Norm Ergodic Theorem of [16] and irreducibility we must have $\check{\mathsf{E}}_x^*[\frac{h_0}{h_*}(\Phi_n)] \to \check{\pi}_*(h_0/h_*)$ as $n \to \infty$ for each x, which gives the first bound.

The second limit involving (g_n) follows from the definition of $\{\widetilde{h}_k\}$ and $\{g_k\}$ which give

$$\log(\widetilde{h}_{n+1}(x)) = \log(V_0(x)) + \sum_{k=0}^{n} [\log(g_k(x)) - \Lambda^*].$$

Dividing by n and using the boundedness of $\{\widetilde{h}_n(x) : n \ge 1\}$ gives the desired limit. □

Lemma 4.4 *Under (A1)–(A4), suppose that*

$$\delta := \inf_{x \in \mathsf{X}} \frac{V_0(x)}{h_*(x)} > 0. \tag{22}$$

Then for all $n \geq 0$ and $x \in \mathsf{X}$,

$$\frac{\widetilde{h}_n(x)}{h_*(x)} \geq \delta.$$

Proof The proof is by induction, where h_0/h_* is bounded from below by $\delta > 0$ by assumption.

If $\widetilde{h}_n/h_*$ is bounded from below by δ then for all x

$$\frac{\widetilde{h}_{n+1}(x)}{h_*(x)} = \frac{c_n P_n \widetilde{h}_n(x)}{\lambda^* h_*(x)} \geq \delta \frac{c_n P_n h_*(x)}{\lambda^* h_*(x)} \geq \delta \frac{c_* P_* h_*(x)}{\lambda^* h_*(x)} = \delta.$$

Hence $\widetilde{h}_{n+1}$ is bounded from below as claimed. □

Theorem 4.5 *Suppose that (A1)–(A4) hold, and suppose that the initial condition satisfies the pair of bounds,*

$$\inf_{x \in \mathsf{X}} \Big(\frac{V_0(x)}{h_*(x)} \Big) > 0; \qquad \check{\pi}_* \Big(\frac{V_0}{h_*} \Big) < \infty.$$

Then $h_n(x) \to h_*(x)$ as $n \to \infty$ for every $x \in \mathsf{X}$.

Proof Let $\check{\mathbf{\Phi}}$ denote the stationary Markov chain with transition probability $\check{P}_*$ and invariant distribution $\check{\pi}_*$. For each $n \leq 0$ we set

$$Z_n = \frac{\widetilde{h}_{-n-1}(\check{\Phi}_n)}{h_*(\check{\Phi}_n)}.$$

From the inequality $\check{P}_* \frac{\widetilde{h}_n}{h_*} \geq \frac{\widetilde{h}_{n+1}}{h_*}$ it follows that $\{(Z_n, \mathcal{F}_n) : n \leq 0\}$ is a submartingale (integrability follows from the bound $\check{\pi}_*(h_0/h_*) < \infty$). Applying [5, Theorem 1, p. 376] we may then conclude that the limit

$$\lim_{n \to -\infty} Z_n = \gamma$$

exists a.s., and since the chain $\check{\mathbf{\Phi}}$ is ergodic, its invariant σ-field is trivial, and hence γ is a constant (c.f. [16, Proposition 17.1.4]). We must also have convergence in probability: For any $\epsilon > 0$, $x \in \mathsf{X}$, as $n \to \infty$,

$$\check{\pi}_*(x) \mathbb{I}\{|\widetilde{h}_{-n-1}(x) - \gamma h_*(x)| > \varepsilon h_*(x)\} = \mathsf{P}\{|Z_n - \gamma| > \varepsilon, \check{\Phi}_n = x\} \to 0,$$

which shows that $\widetilde{h}_n \to \gamma h_*$ pointwise as $n \to \infty$. It follows from Lemma 4.4 that γ is non-zero, and the result then follows since, for each n, the functions h_n and $\widetilde{h}_n$ are constant multiples, and since $h_n(\theta) = h_*(\theta) = 1$ for all n. $\square$

5 Policy Iteration

The policy iteration algorithm, or PIA, is similar to the VIA. Given an initial feedback law w_0 to initialize the algorithm, we denote $\Lambda_0 = \Lambda(w_0, 1)$, so that for any $\theta \in \mathsf{X}$,

$$\mathsf{E}_\theta^{w_0}\Big[\exp\Big(\sum_{k=0}^{\tau_\theta - 1}(C_{w_0}(\Phi_k) - \Lambda_0)\Big)\Big] \le 1. \tag{23}$$

We again recall that the above is an equality provided that $\overline{\alpha}_{w_0} > 1$ (see Theorem 2.2).

One version of the relative value function is given by

$$h_0(x) = \mathsf{E}_x^{w_0}\Big[\exp\Big(\sum_{k=0}^{\sigma_\theta}(C_{w_0}(\Phi_k) - \Lambda_0)\Big)\Big], \qquad x \in \mathsf{X},$$

which satisfies $h_0(\theta) = \exp(C_{w_0}(\theta) - \Lambda_0)$. Provided that w_0 is stabilizing, it follows as in Lemma 2.3 that h_0 is finite valued, uniformly bounded away from zero, and that the multiplicative Poisson inequality holds:

$$P_0 h_0\,(x) \le \lambda_0 c_0^{-1}(x) h_0(x), \qquad x \in \mathsf{X}, \tag{24}$$

where equality holds in (24) provided that (23) is an equality (see [1]).

Given an initial stabilizing feedback law w_0, the PIA defines a sequence of feedback laws, again recursively. Suppose that policies $\{w_0, \ldots, w_n\}$ have been determined together with relative value functions $\{h_0, \ldots, h_n\}$. To enforce the normalization $h_k(\theta) = 1$, we define for all $k \ge 0$, $x \in \mathsf{X}$,

$$h_k(x) := \exp(-C_{w_k}(\theta) + \Lambda_k)\mathsf{E}_x^{w_k}\Big[\exp\Big(\sum_{k=0}^{\sigma_\theta}(C_{w_k}(\Phi_k) - \Lambda_k)\Big)\Big]. \tag{25}$$

A new policy w_{n+1} is then defined to be any solution to the minimization

$$w_{n+1}(x) = \arg\min_{a \in \mathcal{A}} c(x,a) P_a h_n\,(x), \qquad x \in \mathsf{X}. \tag{26}$$

As in the proof of the lower bound on $\{W_N\}$ in Proposition 3.4 we can obtain a uniform lower bound on the relative value functions $\{h_n\}$:

Lemma 5.1 *Suppose that (A1) and (A2) hold. Then there exists $\delta > 0$ such that for each n and x,*

$$h_n(x) \geq \delta > 0$$

□

Like the VIA, the PIA generates stabilizing policies if it is properly initialized:

Theorem 5.2 *Suppose that (A1) and (A2) hold. If w_0 is stabilizing then for any policies $\{w_0, \dots, w_n, \dots\}$ determined by the PIA,*

(i) *Each of the $\{w_0, \dots, w_n, \dots\}$ is stabilizing;*

(ii) *The costs $\{\Lambda_n := \Lambda(w_n, 1), n \geq 0\}$ form a decreasing sequence:*

$$\Lambda_0 \geq \Lambda_1 \geq \cdots \geq \Lambda_n \geq \cdots .$$

Proof The proof is by induction: For any n we have by Lemma 5.1 that $\inf_x h_n(x) > 0$. Also, by minimality,

$$c_{n+1} P_{n+1} h_n(x) \leq c_n(x) P_n h_n(x) \leq \lambda_n h_n(x).$$

From this bound and Lemma 2.3 with $V = h_n$ we conclude that the feedback law w_{n+1} is stabilizing, and $\lambda_n \geq \lambda_{n+1}$. □

Lemma 5.3 *Under (A1) and (A2),*

$$\sup_{n \geq 0} h_n(x) < \infty.$$

Proof Suppose not. Then there exists $x_0 \in \mathsf{X}$, a subsequence $\{n_k\}$ of $\mathbb{Z}_+$, a policy w_∞, and functions h_∞, c_∞ such that as $k \to \infty$,

$$c_{n_k} \to c_\infty, \quad h_{n_k} \to h_\infty, \quad w_{n_k} \to w_\infty,$$

where the convergence is pointwise, and $h_\infty(x_0) = \infty$.

However, from (25) we have $h_\infty(\theta) = 1$, and by Fatou's Lemma,

$$c_\infty(x) P_{w_\infty} h_\infty(x) \leq \lambda_\infty h_\infty(x).$$

Since the control set $\mathcal{A}$ is finite we know that c_∞ is finite valued. It then follows from the above inequality that the set $S = \{x : h_\infty(x) < \infty\}$ is

absorbing. Since it is also non-empty, it must be full [16], and since the kernel P_{w_∞} is irreducible this means that $S = \mathsf{X}$. This is in contradiction to the assumption that $h_\infty(x_0) = \infty$, and we conclude that $\{h_n(x) : n \geq 0\}$ is bounded for any x, as claimed. $\square$

To establish convergence of the PIA to an optimal solution it is necessary to impose some additional assumptions on the process. One convenient assumption is the *skip free* property that for each x, there is a finite set N_x such that $P_a(x, N_x) = 1$, $a \in \mathcal{A}$. This assumption is satisfied for most network models. Unfortunately, we have also been forced to impose some less easily verifiable conditions in Theorem 5.4.

Theorem 5.4 *Suppose that (A1)–(A4) hold; that the kernel P_a is skip free; and suppose that the multiplicative Poisson equation holds for each n:*

$$c_n P_n h_n = \lambda_n h_n.$$

Suppose moreover that

(i) $\check{\pi}_*(\overline{h}/h_*) < \infty$, *where* $\overline{h}(x) = \limsup_n h_n(x)$, $x \in \mathsf{X}$.

(ii) *For any limit $\{w_\infty, h_\infty, c_\infty\}$ of the sequence $\{w_n, h_n, c_n : n \geq 0\}$, the multiplicative Poisson equation has a solution h_{w_∞} for P_{w_∞}; the associated kernel $\check{P}_{w_\infty}$ is positive recurrent with invariant probability $\check{\pi}_{w_\infty}$; and $\check{\pi}_{w_\infty}(h_\infty/h_{w_\infty}) < \infty$.*

Then,

$$\frac{h_n(x)}{h_n(\theta)} \to h_*(x), \qquad x \in \mathsf{X},$$

and $\lambda_n \downarrow \lambda^*$, as $n \to \infty$.

Proof Let $\{\overline{w}, w_\infty, h_\infty, c_\infty\}$ be any subsequential limit of the sequence $\{w_{n+1}, w_n, h_n, c_n : n \geq 0\}$. Clearly $c_\infty = c_{w_\infty}$ and $\lambda_\infty = \inf_n \lambda_n$. By the skip free assumption,

$$\lambda_\infty h_\infty = c_\infty P_{w_\infty} h_\infty.$$

Iterating then gives

$$\begin{aligned}
\frac{h_\infty}{h_{w_\infty}}(x) &= \frac{1}{h_{w_\infty}(x)} \mathsf{E}_x^{w_\infty}\Big[\exp\Big(\sum_{k=0}^{n-1}\Big(C(\Phi_k, w_\infty(\Phi_k)) - \Lambda_\infty\Big)\Big) h_\infty(\Phi_n)\Big] \\
&= \Big(\frac{\lambda(w_\infty)}{\lambda_\infty}\Big)^n \check{\mathsf{E}}_x^{w_\infty}\Big[\frac{h_\infty}{h_{w_\infty}}(\check{\Phi}_n)\Big]
\end{aligned}$$

The expectation on the r.h.s. is bounded due to the ergodicity assumption on P_{w_∞}. We conclude that $\lambda_\infty = \lambda(w_\infty)$ for any limiting feedback law w_∞, and hence also $\lambda_\infty = \lambda(\overline{w})$.

On taking limits we also obtain

$$\lambda_\infty h_\infty = c_\infty P_{w_\infty} h_\infty \geq \overline{c} P_{\overline{w}} h_\infty = \min_w c_w P_w h_\infty .$$

By uniqueness of solutions to the multiplicative Poisson inequality we must have an equality, $\overline{c} P_{\overline{w}} h_\infty = \lambda_\infty h_\infty$ (see Theorem 2.2). That is, $(h_\infty, \overline{w})$ solves the dynamic programming equations for the risk sensitive control problem. Note that this conclusion depends crucially on the observation that $\lambda_\infty = \lambda(\overline{w})$.

Note also that we have not yet shown that $\lambda_\infty = \lambda^*$. For this we iterate the identity

$$c_{w^*} P_* h_\infty \geq \lambda_\infty h_\infty ,$$

to obtain

$$\check{\mathsf{E}}_x^{w^*} \Big[\frac{h_\infty}{h_*} (\check{\Phi}_n) \Big] \geq \Big(\frac{\lambda_\infty}{\lambda^*} \Big)^n \frac{h_\infty}{h_*} (x).$$

Since again the l.h.s. is bounded by assumption, and converges to a limit independent of x, we conclude that $\lambda_\infty = \lambda^*$, and $m := h_\infty / h_*$ is a bounded function of x.

Finally, we have $\check{P}_{w^*} m \geq m$, which shows that m is a bounded, subharmonic function. Since $\check{P}_{w^*}$ is assumed to be recurrent we must have that m is a constant (see [19]), which implies the desired conclusion that $h_\infty(x)/h_\infty(\theta) = h_*(x)$, $x \in \mathsf{X}$. □

6 A queueing model

To illustrate application of the theory we consider an elementary model. Consider the single queue, described by the recursion,

$$Q_{k+1} = [Q_k - u_k + A_{k+1}]_+ , \qquad k \geq 0,$$

where $Q_0 = x \in \mathbb{Z}_+ = \mathsf{X}$ is given. Both $\boldsymbol{Q}$ and $\boldsymbol{u}$ take values in $\mathbb{Z}_+$, and we assume that $u_k \geq 1$ if $Q_k \geq 1$.

Two sources contribute to cost: If Q_k is large then there is excessive inventory, and there is a relatively high price to pay for a large number of servers. With these issues in mind, we take a cost function of the general form,

$$C(x, a) = \theta [g(x) + a], \qquad x \in \mathsf{X},\ a \in \mathbb{Z}_+ ,$$

where $g(x) = o(x)$, so that there is a relatively high cost for servers. We assume that $\theta > 0$, and that $g(x) \to \infty$, $x \to \infty$, so that condition (A1) is satisfied.

The sequence $\{A_k : k \geq 1\}$ is assumed to be i.i.d., and the support of the common marginal-distribution is equal to $\mathbb{Z}_+$. These assumptions imply that the irreducibility condition (A2) holds. The mean of A_1 is necessarily finite, and is denoted α.

Finally, we assume that the moment generating function M_A for A_1 is finite everywhere. This ensures that the risk sensitive cost is finite: To see this, we show that for sufficiently small β, the linear feedback law $w^\circ(x) = \lceil \beta x \rceil$ is stabilizing, where $\lceil z \rceil$ denotes the least integer that is greater than z, $z \in \mathbb{R}$. Consider the Lyapunov function $V_0(x) = e^{\gamma x}$, $x \in \mathsf{X}$, with $\gamma > 0$. We have for any $a \leq x$,

$$\begin{aligned} P_a V_0\,(x) &= \mathsf{E}_x[\exp(\gamma(x - a + A_k))] \\ &= e^{-\gamma a} M_A(\gamma) V_0(x)\,. \end{aligned}$$

Thus, for any $\gamma > \theta$, there exists $\overline{\lambda}_\circ < \infty$ such that

$$P_{w^\circ}\, V_0(x) \leq \overline{\lambda}_\circ \exp(-C_{w^\circ}(x)) V_0(x), \qquad x \in \mathsf{X}\,.$$

The drift inequality (3) holds for this policy, and consequently this linear policy has finite risk-sensitive cost.

An application of Theorem 3.6 shows that an optimal policy w^* exists, with risk sensitive cost $\Lambda^* < \log(\overline{\lambda}_\circ) < \infty$.

Consider now the two algorithms considered above. Theorem 4.5 requires a finite mean $\check{\pi}_*(V_0/h_*)$ to ensure convergence of the VIA. Similar conditions are required in Theorem 5.4 to establish convergence of the PIA.

Suppose that w^* is an optimal policy, and that h_* is the relative value function, so that $P_{w^*} h_* \leq \exp(\Lambda^* - C_{w^*}) h_*$. The following bound is then obtained via Jensen's inequality,

$$P_{w^*} V_*\,(x) \leq \Lambda^* - C_{w^*}(x) + V_*(x) = V_*(x) - \theta[g(x) + w^*(x)] + \Lambda^*$$

where $V_* := \log(h_*)$. Letting τ denote the stopping time,

$$\tau = \min(k : Q_k = 0),$$

we have $Q_\tau = 0$, and we then obtain the bound, for all $x \in \mathsf{X}$,

$$\mathsf{E}_x^{w^*}\Big[\sum_{k=0}^{\tau-1}\Big(\theta[g(Q_k) + w^*(Q_k)] - \Lambda^*\Big)\Big] \leq V_*(x) - V_*(0)\,. \tag{27}$$

We assume without loss of generality that $V_*(0) = \log(h_*(0)) = 0$.

We also have by definition of τ,

$$0 = Q_\tau \geq Q_0 + \sum_{i=0}^{\tau-1} (-w^*(Q_i) + A_{i+1}) \,,$$

from which we deduce that

$$\begin{aligned} x = Q_0 &\leq \mathsf{E}_x^{w^*}\left[\textstyle\sum_{i=0}^{\tau-1}(w^*(Q_i) - A_{i+1}))\right] \\ &= \mathsf{E}_x^{w^*}\left[\textstyle\sum_{i=0}^{\tau-1} w^*(Q_i)\right] - \alpha \mathsf{E}_x^{w^*}[\tau]. \end{aligned}$$

This combined with (27) gives a lower bound on h_*:

$$\log(h_*(x)) \geq \theta(g(x) + x) + (\theta\alpha - \Lambda^*)\mathsf{E}_x^{w^*}[\tau].$$

If $w_*(x) \to \infty$, $x \to \infty$, then $\mathsf{E}_x^{w^*}[\tau] = o(x)$. Unboundedness is a necessary condition for finiteness of the risk sensitive cost when g is unbounded.

For a bound on the mean, note that the lower bound on h_* implies that $V_0(x)/h_*(x) \leq K_\varepsilon \exp(\varepsilon x)$ for any $\varepsilon > \gamma - \theta$, and some finite K_ε. Hence the required bound $\check{\pi}_*(V_0/h_*) < \infty$ is satisfied provided

$$\sum \check{\pi}_*(k) e^{\varepsilon k} < \infty$$

for some $\varepsilon > 0$. This has not been verified, but is plausible when the twisted chain with invariant probability $\check{\pi}_*$ is geometrically ergodic. Hence, given a geometric tail on the steady state distribution $\check{\pi}_*$, it follows that the VIA will converge with the initialization $V_0(x) = e^{\gamma x}$, provided $\gamma > \theta$ is sufficiently small.

References

[1] S. Balaji and S.P. Meyn. Multiplicative ergodic theorems and large deviations for an irreducible Markov chain. *Stochastic Processes and their Applications*, 90(1):123–144, 2000.

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

[3] 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.

[4] R-R. Chen and S.P. Meyn. Value iteration and optimization of multi-class queueing networks. *Queueing Systems*.

[5] Y. Chow and H. Teicher. *Probability Theory: Independence, Interchangeability, Martingales*. Springer-Verlag, New York, NY, 1988.

[6] W. H. Fleming and W. M. McEneaney. Risk sensitive control and differential games. In *Springer Lecture Notes in Control and Info. Sci.*, number 184, pages 185–197. Springer-Verlag, Berlin, Heidelberg, New York, 1992.

[7] 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.

[8] W.H. Flemming and D. Hernández-Hernández. Risk sensitive control of finite state machines on an infinite horizon i. *SIAM J. Control Optim.*, 45:1790–1810, 1997.

[9] P. W. Glynn and S. P. Meyn. A Lyapunov bound for solutions of Poisson's equation. *Ann. Probab.*, 24:916–931, April 1996.

[10] 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.

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

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

[13] M. R. James, J. Baras, and R. J. Elliott. Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems. *IEEE Trans. Automat. Control*, AC-39(4):780–792, April 1994.

[14] Y. Kontoyiannis and S. Meyn. Spectral theory and limit theorems for geometrically ergodic markov processes. Technical report, UIUC, 2000.

[15] G.B.D̃i Masi and L. Stettner. Risk sensitive control of discrete time partially observed Markov processes with infinite horizon. *SIAM J. Control Optim.*, 38(1):61–78, July 1999.

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

[17] S.P. Meyn. The policy improvement algorithm for Markov decision processes with general state space. *IEEE Trans. Automat. Control*, AC-42:1663–1680, 1997.

[18] S.P. Meyn. Algorithms for optimization and stabilization of controlled Markov chains. *SADHANA (Proceedings of the Indian Academy of Sciences, Engineering Sciences)*, 24:339–368, October 1999.

[19] E. Nummelin. *General Irreducible Markov Chains and Nonnegative Operators*. Cambridge University Press, Cambridge, 1984.

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

[21] E. Seneta. *Non-Negative Matrices and Markov Chains*. Springer, New York, NY, 2nd edition, 1981.

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

[23] P. Whittle. *Optimisation: Basics and Beyond*. John Wiley and Sons, Chichester, England, 1996.