Piecewise Stationary Markov Decision Processes, II: State-Dependent Gain

M. Jacobson, N. Shimkin and A. Shwartz

Department of Electrical Engineering
Technion, Israel Institute of Technology
Haifa 32000, Israel

November 10, 1999
Submitted to MOR

Abstract

This paper continues the study of Piecewise Stationary Markov Decision Processes (PSMDPs) with discounted costs. In such models, each decision cycle is partitioned into $N+1$ *fastscale* epochs. In the first N epochs the process evolves according to an *underlying MDP* with stationary data, while at the final epoch the rewards and transition law are distinct. For N large, *initially stationary policies* are natural candidates for optimal policies. In this paper, we consider the general case where the underlying MDP has a multi-chain structure, hence a state-dependent optimal average reward (gain). Results are developed describing the form of the PSMDP's optimal value vector as a function of N. We also give conditions under which $N\epsilon$-optimal initially stationary policies exist. These policies require a planning horizon length with a bound depending only on ϵ and not on N. Under somewhat stronger conditions, ϵ-optimal initially stationary policies exist. While the planning horizon length may increase as a function of N in this case, it is shown to be of lower order than N asymptotically. Counter-examples are presented which demonstrate that these results may fail to hold if our conditions are relaxed.

1 Introduction

The work presented here is a sequel to [3]. There, we introduced the notion of a Piecewise Stationary Markov Decision Process (PSMDP). In these models, the decision horizon is partitioned into intervals of $N+1$ epochs which we call *renewal cycles.* In the first N epochs of each renewal cycle (called the *stationary epochs*), the process evolves according to a time-homogeneous set of transition probability and reward functions. However, at the final epoch (the *non-stationary* *epoch*), the rewards and transition functions are distinct. These distinct functions capture the effect of decisions made at a slower time scale.

In [3], we considered a discounted reward optimality criterion in which rewards devalue at the start of each renewal cycle. We analyzed the solutions of the optimality problem for the case where the average reward MDP associated with the stationary data (called the underlying MDP) had an optimal gain which was independent of the initial state. In what follows, we consider the case where the gain may be state-dependent. An illustrative application where state-dependent gain is naturally encountered is the *alternative projects management* problem. In such applications, a decision maker works on one of several projects for a period of N epochs. At the end of each period, she may switch to another project. These applications may be modeled as PSMDPs whose underlying MDP is multichain with subchains corresponding to the different projects.

Our aim is to examine whether initially stationary policies are approximately optimal. In addition, we examine whether the required planning horizon length is small compared to N. Though some review is given, for a full description and discussion of the model we refer to [3]. In the state-independent gain case, it was found, under fairly weak conditions on the underlying MDP, that ϵ-optimal initially stationary policies (i.s.p.'s) exist. The required planning horizon η_ϵ was found to have a bound depending only on ϵ and not on N. Moreover, the initial decision rules were gain optimal in the underlying MDP. This was because certain relevant value iteration sequences converged in a bounded number of steps.

Conversely, in the state-dependent gain case, these sequences do not generally converge in a bounded number of steps. Examples 8.3 and 8.4 in [3] illustrated the type of behavior which may be observed as a result. In Example 8.3, all ϵ-optimal policies had horizons which were unbounded. In Example 8.4, the only (ϵ-)optimal i.s.p. in the model had a initial decision rule which was not gain optimal.

The analysis presented in this paper identifies conditions on the underlying MDP that ensure the existence of ϵ-optimal i.s.p.'s with *diminishing* planning horizons. We call a planning horizon diminishing if it is given by a function which is of lower order than N asymptotically. Furthermore, it is established that $N\epsilon$-optimal i.s.p.'s exist if N is large enough. Since the opportunity for rewards grows linearly in N in the state-dependent gain case (see Theorem 8.1(b) in [3]), there is adequate justification in accepting this kind of "scaled ϵ-optimality".

The conditions upon which these results rely are discussed in Section 2.4. Essentially, these conditions impose requirements on the the chain structure of the underlying MDP. They are naturally satisfied in the alternative project management applications in which the projects may be individually modeled as weakly communicating MDPs.

This paper is organized as follows. In Section 2, we review the PSMDP model, supplement the notation of [3], and describe the assumptions used in our analysis. In Section 3, we establish some preliminary results.

In Section 4, we analyze the form of the optimal value as a function of N. We find that it may be decomposed into the sum of a term which is linear in N and one which is of $O(\log N)$. Unlike the state-independent gain case, the linear term in the decomposition depends on both the stationary and non-stationary data.

In Section 5, we present Theorem 5.1 which establishes that, under appropriate assumptions, there exists a uniform $N\epsilon$-optimal i.s.p. whose planning horizon, η_ϵ, is a function only of ϵ. Moreover, the initial decision rule may be derived from the underlying MDP, Ψ, alone. Theorem 5.2, indicates that a similar i.s.p. exists whose planning horizon $\eta(N)$ is of order $\log N$. However, the i.s.p. is guaranteed to be optimal up to a term which is only $O(\log N)$. So, at the price of a slightly larger planning horizon, one can approximate optimality much better.

In Section 6, we establish conditions for the existence of uniform ϵ-optimal initially stationary policies having diminishing planning horizons. The proof shows that the number of steps required for the relevant non-discounted value iteration sequences to approximately converge is of lower order than N. In the state-dependent gain case, there are two obstacles in this proof. Firstly, the value iteration sequences are not guaranteed to reach their geometric phase of convergence in a number of steps which is less than N. Secondly, even when this phase is reached, the geometric convergence rate is generally not uniform over an unbounded set of terminal reward vectors. These two obstacles are surmounted with the help of some new results on non-discounted value iteration.

Finally, in Section 7, we present a counter-example to the results of the paper when the conditions hypothesized by them do not hold.

2 Model Description and Notations

This section describes the model, notation, and assumptions which we work with throughout. Our analysis focuses on Delayed Slowscale Models (DSMs). Extensions to Markovian Slowscale Models (MSMs), as discussed in [3], are readily obtainable from results for DSMs in which A^σ is a singleton.

2.1 Review of the Model and Basic Notation

We briefly review here some model definitions and notation previously introduced in [3]. A Piecewise Stationary Markov Decision Process (PSMDP) is a discrete-time sequential decision process whose evolution is described by an $(N+1)$-periodic sequence of transition probability and reward functions. In the first N epochs of each periodic interval, the probability and reward functions are time-homogeneous. At the final epoch, however, they are distinct. We call these periodic intervals *renewal cycles* and the epochs at the start of these cycles – i.e. epochs $t = 0, (N+1), 2(N+1), \dots$ – *renewal epochs*. The first N epochs of each renewal cycle are called *stationary epochs* and the reward and transition probability functions governing the stationary epochs are referred to as the *stationary data*. The final epoch in each renewal cycle is called the *non-stationary epoch* and the relevant reward and transition probability functions are referred to as the *non-stationary data*.

A Delayed Slowscale Model (DSM) is a PSMDP specified by a state space S, a space of allowable fastscale actions A_s for each state $s \in S$, and a space of slowscale actions A^σ. Fastscale actions, $a_t \in A_s$, are selected at all epochs t while slowscale actions $a_t^\sigma \in A^\sigma$ are selected at renewal epochs. Rewards and transition probabilities are given at stationary epochs by functions $r(s_t, a_t)$ and $p(s_{t+1}|s_t, a_t)$. However, at non-stationary epochs, they are given as $r(s_t, a_t, a^\sigma_{(N+1)\lfloor t/(N+1)\rfloor})$, $p(s_{t+1}|s_t, a_t, a^\sigma_{(N+1)\lfloor t/(N+1)\rfloor})$. We assume finite state and action spaces throughout.

We represent functions on S by column vectors in $\mathbb{R}^{|S|}$. For a given $y \in \mathbb{R}^{|S|}$, $||y||$ denotes the sup-norm and $||y||_{\text{sp}}$ denotes the span semi-norm. For a scalar $c \geq 0$, $\bar{o}(c)$ denotes an unspecified vector in $\mathbb{R}^{|S|}$ whose sup-norm is at most c.

The basic dynamic programming operators, written here in vector form, are

$$\begin{aligned} Lx &\triangleq \max_{d \in D}\{r_d + P_d x\} \\ L^{a^\sigma} x &\triangleq \max_{d \in D}\{r_d^{a^\sigma} + \lambda P_d^{a^\sigma} x\}, \quad a^\sigma \in A^\sigma \end{aligned}$$

where D is the space of possible decision rules mapping each $s \in S$ to an $a^\sigma \in A_s^\sigma$.

For a decision rule $d \in D$, $L_d = r_d + P_d$ represents the restriction of L to d. If $\pi = \{d_m, d_{m-1}, \dots, d_1\}$ represents a sequence of m decision rules, then for all $0 \leq k \leq m$,

$$L_\pi^k x \triangleq r_{d_k} + P_{d_k} r_{d_{k-1}} + P_{d_k} P_{d_{k-1}} r_{d_{k-2}} + \dots + (P_{d_k} P_{d_{k-1}} \cdots P_{d_1}) x.$$

Analogous notation is defined for other one-step dynamic programming operators.

The DSM discounted dynamic programming operator is

$$\mathcal{L}_N x(s) = \max_{a^\sigma \in A^\sigma}\{L^N L^{a^\sigma} x(s)\}, \quad s \in S \ . \tag{2.1}$$

With this notation, the DSM optimality equation can be written $v_N^* = \mathcal{L}_N v_N^*$ or, component-wise,

$$v_N^*(s) = \max_{a^\sigma \in A^\sigma}\{L^N L^{a^\sigma} v_N^*(s)\}, \quad s \in S \ . \tag{2.2}$$

The maximizations on the right hand side of (2.1) are obtained by sequences of the form

$$\{a^\sigma, d_0, d_1, \dots, d_{N-1}, d\}(s), \quad s \in S \ . \tag{2.3}$$

Letting $\pi = \{d_0, d_1, \dots, d_{N-1}\}$, this abbreviates to $\{a^\sigma, \pi, d\}$. The sequences of the form (2.3) specify a cyclo-stationary policy. At each renewal epoch, a cyclo-stationary policy prescribes, based on the current state s, a slowscale action a^σ to be taken immediately as well as a sequence of decision rules for choosing fastscale actions throughout the upcoming renewal cycle.

An initially stationary policy (i.s.p.) is a cyclo-stationary policy which prescribes a sequence of the form

$$\{a^\sigma, \delta, \delta, \delta, \ldots, \delta, d_{N-\eta}, d_{N-\eta+1}, \ldots, d_N\}(s).$$

The decision rule δ is called the initial decision rule for s. The parameter η is called the *planning horizon.* When the initial decision rule is the same for every state s, the i.s.p. is called *simple.*

We shall refer to a planning horizon as *diminishing* if it is given by a non-negative, integer function $\eta(N)$ which is of lower order than N asymptotically, i.e. $\eta(N)/N \to 0$. In general, we shall refer to functions of lower order than N as diminishing. Likewise, a quantity shall be referred to as diminishing if it given by a diminishing function.

2.2 Properties of the Underlying MDP

The stationary data can be associated with an average reward MDP, denoted as Ψ, with decision rule space D and optimal gain vector g^*. We call Ψ the *underlying MDP.* Here, we define some notation for Ψ and recall some relevant properties of average reward MDPs (see also [5]).

Define

$$\begin{aligned} Ux &\stackrel{\triangle}{=} \max_{d \in D}\{P_d x\} \\ Tx &\stackrel{\triangle}{=} \max_{d \in E}\{r_d + P_d x\} \end{aligned}$$

where $E \stackrel{\triangle}{=} \{d \in D \mid P_d g^* = g^*\}$. The optimality equations of Ψ can then be written,

$$g^* = Ug^* \tag{2.4}$$

$$g^* + v = Tv \tag{2.5}$$

where v belongs to a closed, unbounded set of solutions V.

If P_d^* represents the limiting matrix of P_d, then $g_d = P_d^* r_d$ denotes the gain of d. The set $D^* \stackrel{\triangle}{=} \{d \in D \mid g^* = g_d\}$ is the set of optimal decision rules in Ψ. By Theorem 3.1(e) in [5], the necessary and sufficient conditions for a decision rule d (randomized or deterministic) to be average optimal are that, for any fixed $v \in V$,

1. $P_d g^* = g^*$.

2. $g^*(s) + v(s) = r_d(s) + P_d v(s)$ for all $s \in S$ which are recurrent states of P_d.

An immediate consequence of requirement 1 is that $D^* \subset E$.

Let $E(v) \stackrel{\triangle}{=} \{d \in E \mid g^* + v = r_d + P_d v\}$. The set V is known to be convex if and only if there exists a $d \in D$ such that $d \in E(v)$ for all $v \in V$. This follows from Theorem 4.3 and Theorem 3.2(g) in [5].

Define

$$R^* \stackrel{\triangle}{=} \{s \in S \mid s \text{ is recurrent for some } \textit{average optimal} \text{ decision rule}\}.$$

In [5], it is shown that R^* has a unique decomposition,

$$R^* = \bigcup_{\alpha=1}^{n^*} R^*(\alpha).$$

The sets $R^*(\alpha), \alpha = 1, \ldots, n^*$ which, for convenience, we refer to as the *Schweitzer-Federgruen classes*, are mutually disjoint with the following properties, (see Theorem 3.2 of [5]):

1. Any irreducible subchain of any optimal randomized decision rule is contained in one of the sets $R^*(\alpha)$.
2. For each $\alpha = 1, \ldots, n^*$, a randomized optimal decision rule exists which has $R^*(\alpha)$ as an irreducible subchain, i.e. $R^*(\alpha)$ is a communicating set.

Finite algorithms are known for identifying the Schweitzer-Federgruen classes (see, for example, the discussion in [5, pages 314]). It is known (cf. Theorem 5.1 in [5]) that, for any given $v_1, v_2 \in V$, the difference $v_1(s) - v_2(s)$ is constant as a function of s over any fixed $R^*(\alpha)$.

2.3 Zero-Reward Analogues

When we consider a version of Ψ in which all rewards are set to zero, we obtain useful analogues to the properties cited in the preceding Section. Firstly, equations (2.4) and (2.5) reduce to

$$w = \max_{d \in D}\{P_d w\} = Uw \tag{2.6}$$

where w has replaced v.

The set $W \stackrel{\triangle}{=} \{w \in R^{|S|} \mid w = Uw\}$ is the analogue of V. Observe that the vectors $\mathbf{1}$ and g^* are in W. In addition, since U is *positive homogeneous*, i.e.

$$Ucx = cUx$$

for any non-negative scalar c, it follows that W is a cone.

Define

$$K(w) \triangleq \{d \in D \mid P_d w = Uw = w\}$$

as the set of decision rules attaining the maximum in (2.6). In analogy with V, the set W is convex if and only if there exists a decision rule $d \in D$ such that $d \in K(w)$ for all $w \in W$. Note, from the positive homogeneity of U, that $K(cw) = K(w)$ for any non-negative scalar c.

In addition, we define, for all $w \in W$,

$$\Delta w \triangleq \begin{cases} \min\limits_{d \in D \setminus K(w)} ||Uw - P_d w|| & : K(w) \neq D \\ \infty & : K(w) = D. \end{cases}$$

Clearly, $\Delta w > 0$ and $\Delta w = \infty$ if and only if $K(w) = D$. Moreover,

$$\operatorname*{argmax}_{d \in D}\{P_d w + \frac{1}{2}\bar{o}(\Delta w)\} = \operatorname*{argmax}_{d \in K(w)}\{P_d w + \frac{1}{2}\bar{o}(\Delta w)\}. \tag{2.7}$$

Since all decision rules are optimal in the zero-reward version, we have the following analogue

$$\hat{R} \triangleq \{s \in S \mid s \text{ is recurrent for some decision rule}\}$$

of the set R^*. Likewise, we have the following analogous decomposition, sometimes referred to as the Bather decomposition [1].

$$\hat{R} = \bigcup_{\alpha=1}^{\alpha^*} \hat{R}(\alpha).$$

This decomposition has the following properties

1. Any irreducible subchain of any randomized decision rule is contained in one of the sets $\hat{R}(\alpha)$.

2. For each $\alpha = 1, \ldots, \alpha^*$, a randomized decision rule exists which has $\hat{R}(\alpha)$ as a subchain, i.e. $\hat{R}(\alpha)$ is a communicating set.

We will refer to the sets $\hat{R}(\alpha)$ as the *Bather classes.* Finite algorithms are known for identifying the Bather classes (see, for example, [8]). When there is only one Bather class, Ψ is weakly communicating.

By analogy with V, the difference between elements of W is state-independent over each fixed $\hat{R}(\alpha)$. Since $\mathbf{1} \in W$, it follows that all $w \in W$ are state-independent on the Bather classes.

2.4 Assumptions

We shall work with combinations of the following conditions on the underlying MDP, Ψ. The statement of condition (C3) here is equivalent to that in [3].

(C3) For all terminal rewards x, the sequence $T^k x - kg^*$ converges.

(C4) For all terminal rewards x, the sequence $U^k x$ converges.

(C5) There exists a deterministic average optimal decision rule, $\gamma \in D^*$, satisfying, for all $w \in W$

$$P_\gamma w = w = Uw. \tag{2.8}$$

Hence, $\gamma \in K(w)$ for all $w \in W$.

Conditions (C3) and (C4) are equivalent to non-periodicity requirements on the chain structure of Ψ. Condition (C3) is equivalent [6] to the condition that a randomized, aperiodic, gain optimal decision rule exists whose recurrent chains are the Schweitzer-Federgruen classes $R^*(\alpha)$, $\alpha = 1, \dots, n^*$. By analogy with an MDP with zero rewards, (C4) is equivalent to the condition that a randomized, aperiodic decision rule (not necessarily gain optimal) exists whose subchains are $\hat{R}(\alpha)$, $\alpha = 1, \dots, \alpha^*$.

It is often complicated to verify conditions (C3) or (C4) directly. A convenient condition guaranteeing that both assumptions hold is that all deterministic decision rules in Ψ induce aperiodic chains.

Assumption (C5) is a condition which, to our knowledge, has not been considered before. It requires the existence of a decision rule $\gamma \in K(w)$ for all w, or equivalently that W be convex (see Section 2.3). In addition, it requires that γ be gain optimal.

A sufficient condition for (C5) is that the following two hypotheses hold.

(H1) There are no states which are transient under all policies $d \in D$, i.e. $S \setminus \hat{R} = \{\emptyset\}$.

(H2) There is a Schweitzer-Federgruen class in each Bather class, i.e. $\hat{R}(\alpha) \cap R^* \neq \{\emptyset\}$, $\alpha = 1, \dots, \alpha^*$.

When (H1) holds, the Bather classes completely partition the state space, S. When (H2) holds, an average optimal decision rule exists under which all the Bather classes are closed. Denoting this decision rule γ, it follows from the fact that all vectors, $w \in W$, are constant over the Bather classes, that

$$P_\gamma w = w = Uw.$$

which is (2.8).

Another hypothesis which implies (C5) is

(H3) The state space has a partition into subchains each containing a single Bather class and which are globally closed, i.e. they do not communicate under any decision rule.

Hypothesis (H3) implies (C5) because all elements of W are then state-independent over each globally closed block. This is because each block can be regarded as a separate, weakly communicating sub-MDP. The latter condition is the zero-reward analogue of (C2) in [3]. Hence, all $w \in W$ differ from $\mathbf{1}$ by a state-independent vector on each block.

Hypothesis (H3) is a condition which might naturally be encountered in alternative projects models. In such models, an underlying MDP exists with a partition into globally closed blocks, each one representing a project. The decision maker works on one of these projects for intervals of N epochs and then, at non-stationary epochs, may select a new project. Each project can be modeled as a separate MDP. If each MDP is weakly communicating, then (H3) holds.

Remark 2.1 Hypothesis (H2) holds trivially when there is only one Bather class. It is also valid when there are two Bather classes and the values of g^* corresponding to each Bather class are distinct. In this case, all average optimal decision rules have a recurrent class in each Bather class.

To see this, suppose, by way of contradiction that, for some $d \in D^*$, all states in Bather class 1 are transient. Then via P_d^*, only Bather class 2 is reached with positive probability. This implies that $P_d^* g^* = g_2^* \mathbf{1} \neq g^*$ where g_2^* is the value of g^* on Bather class 2. However, this is a contradiction, because all average optimal decision rules, d, satisfy $P_d^* g^* = g^*$.

Remark 2.2 Hypothesis (H2), but not (H1), is implied by (H3).

2.5 Additional Notation

When (C3) holds, the operator

$$\hat{T}^{\infty} x \triangleq \lim_{k \to \infty} T^k x - k g^*$$

is defined for all $x \in \mathbb{R}^{|S|}$. As established in [7], $||T^k x - k g^* - \hat{T}^{\infty} x||$ converges monotonically to zero at a geometric rate which depends on x. Also, $\hat{T}^{\infty}$ maps into V and has all the common properties of a dynamic programming operator.

Analogously, when (C4) holds, the operator

$$\hat{U}^{\infty} x \triangleq \lim_{k \to \infty} U^k x. \tag{2.9}$$

is defined for all $x \in \mathbb{R}^{|S|}$. Convergence of $||U^k x - \hat{U}^\infty x||$ to zero is monotonic and geometric with a rate which depends on x. The $\hat{U}^\infty$ operator maps into W and has all the same properties as the operator U.

Now, let us denote

$$U^{a^\sigma} x \triangleq \max_{d \in D}\{P_d^{a^\sigma} x\}.$$

as the analogue of U for the non-stationary data. When (C4) holds, the operator

$$Qx(s) \triangleq \max_{a^\sigma \in A^\sigma}\{g^*(s) + \lambda \hat{U}^\infty U^{a^\sigma} x(s)\}, \quad s \in S$$

is a contraction mapping on $\mathbb{R}^{|S|}$ with respect to $||\cdot||$ with contraction factor λ. We denote it's fixed point by x_∞.

For each $a^\sigma \in A^\sigma$, we define the restricted decision rule sets

$$D^{a^\sigma} \triangleq E \cap K(\hat{U}^\infty U^{a^\sigma} \lambda x_\infty).$$

and the MDP, Ψ^{a^σ}, as the restriction of Ψ whose decision rule set is D^{a^σ}. When (C5) holds, each D^{a^σ} is non-empty (since clearly $\gamma \in D^{a^\sigma}$). Similarly, let us define

$$T_{(a^\sigma)} x \triangleq \max_{d \in D^{a^\sigma}}\{r_d + P_d x\}.$$

We see immediately that $T_{(a^\sigma)}$ is the L operator for Ψ^{a^σ}. What may be less obvious is that $T_{(a^\sigma)}$ is also the T operator for Ψ^{a^σ}. Because γ is optimal in Ψ and $\gamma \in D^{a^\sigma}$, clearly g^* must also be the optimal gain of Ψ^{a^σ}. Moreover, for any $d \in D^{a^\sigma}$,

$$P_d g^* = g^*$$

since, by the definition of D^{a^σ}, d is also in E. Therefore, all decision rules in D^{a^σ} are maximizing in the first optimality equation of Ψ^{a^σ}. Hence, $T_{(a^\sigma)}$ is this model's T operator analogue. It follows that, under appropriate conditions, the sequence

$$T^n_{(a^\sigma)} x - n g^*$$

will exhibit all of the convergence characteristics classically associated with the T operator.

For every finite set of reward vectors G, and every $\epsilon > 0$, let

$$M_U(G, \epsilon) \triangleq \min\{k \; : \; ||U^k x - \hat{U}^\infty x|| \le \epsilon \text{ for all } x \in G\}.$$

Since $||U^k x - \hat{U}^\infty x||$ converges monotonically, $M_U(G, \epsilon)$ is the point at which all the sequences $U^k x$, $x \in G$ have converged within ϵ.

In general, we will sometimes refer to the number of steps for a backward induction sequence to converge within a certain accuracy as its "convergence time". The function M_U gives the maximum convergence time for $U^k x$, $x \in G$.

The set

$$G_{x_\infty} \triangleq \{x \in \mathbb{R}^{|S|} \mid x = U^{a^\sigma} \lambda x_\infty \text{ for some } a^\sigma \in A^\sigma\}$$

will be of particular interest, as will

$$M_{x_\infty}(\epsilon) \triangleq M_U(G_{x_\infty}, \epsilon).$$

Finally, we define $\bar{r}$ as a bound on all of the reward data in the PSMDP.

3 Preliminary Results

In this section, we derive some results pertaining to the dynamic programming operators L, U, and $\mathcal{L}_N$ which will assist us in the main part of our analysis. The first Lemma establishes some useful properties concerning W – the set of fixed points of U – in the case when it is convex. Recall from Section 2.4 that this is always the case when (C5) holds.

Lemma 3.1 (Convexity properties of W) *Suppose W is convex. Let $w_1, w_2 \in W$ and $c_1, c_2 > 0$. Then*

(a) $c_1 w_1 + c_2 w_2 \in W$.

(b) $K(c_1 w_1 + c_2 w_2) = K(w_1) \cap K(w_2)$.

(c) $\Delta(c_1 w_1 + c_2 w_2) \geq \min(c_1, c_2) \min(\Delta w_1, \Delta w_2)$.

Proof.

(a) Immediate from the fact that W is a convex cone.

(b) We first fix $d \in K(c_1 w_1 + c_2 w_2)$, implying

$$\begin{aligned} c_1 w_1 + c_2 w_2 &= U(c_1 w_1 + c_2 w_2) \\ &= c_1 P_d w_1 + c_2 P_d w_2. \end{aligned} \tag{3.1}$$

Let us further suppose that $d \notin K(w_1) \cap K(w_2)$. Then at least one of $P_d w_1 < w_1$ or $P_d w_2 < w_2$ is true. Since $c_1, c_2 > 0$,

$$c_1 w_1 + c_2 w_2 < c_1 w_1 + c_2 w_2$$

is obtained from (3.1), establishing a contradiction.

This shows that that $K(c_1w_1+c_2w_2) \subseteq K(w_1) \cap K(w_2)$. To demonstrate the reversed inclusion, fix $d \in K(w_1) \cap K(w_2)$. Then

$$\begin{aligned} U(c_1w_1 + c_2w_2) &\geq c_1P_dw_1 + c_2P_dw_2 \\ &= c_1w_1 + c_2w_2 \\ &= U(c_1w_1 + c_2w_2). \end{aligned}$$

where the last equality followed from part (a). Hence, $P_d(c_1w_1 + c_2w_2) = U(c_1w_1 + c_2w_2)$ implying that $d \in K(c_1w_1 + c_2w_2)$.

(c) Since $c_1w_1, c_2w_2 \in W$ and since $\min(\Delta c_1w_1, \Delta c_2w_2) \geq \min(c_1, c_2)\min(\Delta w_1, \Delta w_2)$, it is sufficient to prove (c) for $c_1 = c_2 = 1$.

Let $w_3 = w_1 + w_2$. For any $d \in D$,

$$(I - P_d)w_3 = (I - P_d)w_1 + (I - P_d)w_2. \tag{3.2}$$

Since both w_1 and w_2 are in W, both $(I - P_d)w_1$ and $(I - P_d)w_2$ are non-negative. Furthermore, $w_i = Uw_i$ for $i = 1, 2, 3$. Equation (3.2) therefore implies

$$||Uw_3 - P_dw_3|| \geq ||Uw_1 - P_dw_1||, \tag{3.3}$$

$$||Uw_3 - P_dw_3|| \geq ||Uw_2 - P_dw_2||. \tag{3.4}$$

Now, from part (b), $K(w_3) = K(w_1) \cap K(w_2)$. If $K(w_3) = D$, then $\Delta w_3 = \infty$ and the result follows trivially. If this is not the case, then

$$\Delta w_3 = \min_{d \in K(w_1)^c \cup K(w_2)^c} ||Uw_3 - P_dw_3||.$$

Here, $K(w)^c$ denotes the set complement of $K(w)$. If the minimum in the above is achieved in $K(w_1)^c$, (3.3) implies

$$\begin{aligned} \Delta w_3 &= \min_{d \in K(w_1)^c} ||Uw_3 - P_dw_3|| \\ &\geq \min_{d \in K(w_1)^c} ||Uw_1 - P_dw_1|| \\ &= \Delta w_1 \\ &\geq \min(\Delta w_1, \Delta w_2), \end{aligned}$$

whereas if the minimum is achieved in $K(w_2)^c$, (3.4) similarly implies

$$\begin{aligned} \Delta w_3 &\geq \Delta w_2 \\ &\geq \min(\Delta w_1, \Delta w_2). \end{aligned}$$

Since one of these two conditions holds, part (c) follows. $\square$

The following Lemma gives us a logarithmic bound on $M_U(G, \epsilon)$.

Lemma 3.2 (Log bound on number of steps to convergence) *Suppose (C4) holds and consider any finite set of vectors, G. Then there exist constants $J, b_G > 0, k_G > 1$ such that $M_U(G, \epsilon) \leq \max(J \left\lceil \log_{k_G} \left(\frac{b_G}{\epsilon} \right) \right\rceil, 0)$.*

Proof.

Immediate from the finiteness of G and the geometric rate of convergence of $U^k(\cdot)$. □

The following Proposition describes value iteration on terminal reward vectors of the form Nx_0.

Proposition 3.3 (Value iteration on scaled vectors) *Assume (C4) and (C5). Let y_0 be a terminal reward vector of the form $y_0 = Nx_0$, let G be a finite set of vectors containing x_0, and fix any $\epsilon > 0$.*

Then, the representation,

$$L^k y_0 = kg^* + N(\hat{U}^\infty x_0 + \bar{o}(\epsilon)) + \bar{o}(c_1 M_U(G, \epsilon)) + \bar{o}(c_2)$$

holds for all $k \geq M_U(G, \epsilon)$, where c_1, c_2 are positive constants.

Proof.

For brevity, let $M = M_U(G, \epsilon)$. We denote a generic M-horizon policy as $\pi = \{d_1, d_2, \ldots, d_M\}$ and the M-step transition matrix it induces as P_π^M. Then,

$$\begin{aligned}
L^M y_0 &= \max_\pi \{r_{d_1} + P_{d_1} r_{d_2} + \cdots + P_\pi^M (Nx_0)\} \\
&= N \max_\pi \{P_\pi^M x_0\} + \bar{o}(\bar{r}M) \\
&= NU^M x_0 + \bar{o}(\bar{r}M). \qquad (3.5)
\end{aligned}$$

However, by the definition of M, we have

$$U^M x_0 = \hat{U}^\infty x_0 + \bar{o}(\epsilon)$$

so that (3.5) becomes

$$L^M y_0 = N(\hat{U}^\infty x_0 + \bar{o}(\epsilon)) + \bar{o}(\bar{r}M). \qquad (3.6)$$

Now from the proof in Theorem 9.4.1 in [4] of the boundedness of $L^n x - ng^*$ we know that for a certain vector h,

$$\begin{aligned}
L^n x &= r_{d_1} + P_{d_1} r_{d_2} + \cdots + P_\pi^n x \\
&\leq ng^* + h + P_\pi^n (x - h) \qquad (3.7)
\end{aligned}$$

for any given terminal reward vector x. In addition, if we consider the 'MDP' consisting of a single average optimal decision rule $\delta \in D^*$, we may also derive from the proof the lower bound,

$$\begin{aligned} L^n x &\geq L^n_\delta x \\ &= ng^* + h_\delta + P^n_\delta (x - h_\delta). \end{aligned} \tag{3.8}$$

where h_δ is the bias vector of δ.

Taking $x = L^M y_0$ and $n = k - M$, (3.7) becomes

$$\begin{aligned} L^{k-M} L^M y_0 &= L^k y_0 \\ &\leq (k - M) g^* + h + P^n_\pi L^M y_0 - P^n_\pi h. \end{aligned}$$

Upon substituting (3.6), rearranging, and noting that $\hat{U}^\infty x_0 \in W$,

$$\begin{aligned} L^k y_0 &\leq kg^* + N(P^n_\pi \hat{U}^\infty x_0 + \bar{o}(\epsilon)) + \bar{o}(2||h||) + \bar{o}(2\bar{r}M) \\ &\leq kg^* + N(\hat{U}^\infty x_0 + \bar{o}(\epsilon)) + \bar{o}(2||h||) + \bar{o}(2\bar{r}M). \end{aligned} \tag{3.9}$$

Similarly, (3.8) becomes, once we let $\delta = \gamma$, where γ is as in (C5)

$$\begin{aligned} L^k y_0 &\geq kg^* + N(P^n_\gamma \hat{U}^\infty x_0 + \bar{o}(\epsilon)) + \bar{o}(2||h_\gamma||) + \bar{o}(2\bar{r}M) \\ &= kg^* + N(\hat{U}^\infty x_0 + \bar{o}(\epsilon)) + \bar{o}(2||h_\gamma||) + \bar{o}(2\bar{r}M). \end{aligned} \tag{3.10}$$

From (3.9) and (3.10), we deduce that

$$L^k y_0 - (kg^* + N\hat{U}^\infty x_0) = N\bar{o}(\epsilon) + \bar{o}(c_1 M) + \bar{o}(c_2)$$

Here, we have taken c_2 as a bound on the terms $\bar{o}(2||h||), \bar{o}(2||h_\gamma||)$ and also replaced $2\bar{r}$ by c_1. The conclusions of the Theorem are therefore proven. □

The following Proposition relates a single step of value iteration $\mathcal{L}_N(Nx_0)$ to Q.

Proposition 3.4 (Two timescale value iteration on scaled vectors) *Assume (C4) and (C5). Let z_0 be a terminal reward vector of the form $z_0 = Nx_0$, define $G = \{U^{a^\sigma} \lambda x_0, a^\sigma \in A^\sigma\}$ and fix any $\epsilon > 0$.*

Then, if $N \geq M_U(G, \epsilon)$, the representation

$$\mathcal{L}_N z_0 = N(Qx_0 + \bar{o}(\epsilon)) + \bar{o}(c_1 M_U(G, \epsilon)) + \bar{o}(c_2) \tag{3.11}$$

holds for positive constants c_1, c_2.

Proof.

Fix $a^\sigma \in A^\sigma$ and let $x^{a^\sigma} = U^{a^\sigma}\lambda x_0$. Then,

$$\begin{aligned} L^{a^\sigma} z_0 &= \max_{d\in D}\{r_d^{a^\sigma} + \lambda P_d^{a^\sigma}(Nx_0)\} \\ &= N\max_{d\in D}\{P_d^{a^\sigma}\lambda x_0\} + \bar{o}(\bar{r}) \\ &= NU^{a^\sigma}\lambda x_0 + \bar{o}(\bar{r}) \end{aligned}$$

By the nonexpansive property of L^N, therefore,

$$L^N L^{a^\sigma} z_0 = L^N(NU^{a^\sigma}\lambda x_0) + \bar{o}(\bar{r}). \tag{3.12}$$

Because $U^{a^\sigma}\lambda x_0 \in G$, the hypotheses of Proposition 3.3 are satisfied for an initial reward vector of the form $NU^{a^\sigma}\lambda x_0$ and $k = N$. Hence, it applies to the first term on the on the right- hand side, yielding

$$L^N L^{a^\sigma} z_0 = Ng^* + N(\hat{U}^\infty U^{a^\sigma}\lambda x_0 + \bar{o}(\epsilon)) + \bar{o}(c_1 M_U(G,\epsilon)) + \bar{o}(\tilde{c}_2).$$

Since a^σ was arbitrary, we can maximize both sides of the last equation over A^σ to obtain (3.11).
□

4 Asymptotic Behavior of v_N^*

In this section, we examine the N-dependence of v_N^*. The next Theorem is the main result of this section. It shows that under (C4) and (C5), the optimal value v_N^* can be decomposed into an explicit term which is linear in N and a term which is $O(\log N)$. In contrast to Theorem 4.1 in [3], we find that in the state-dependent gain case, the linear term may depend on both the stationary and non-stationary data.

Theorem 4.1 (Asymptotic behavior of v_N^* - non-constant gain) *Assume (C4) and (C5). Then*

$$v_N^* = Nx_\infty + O(\log N). \tag{4.1}$$

Proof.

Let $z_0 = Nx_\infty$ and let $G = \{U^{a^\sigma}\lambda x_\infty, a^\sigma \in A^\sigma\}$. Fix ϵ and suppose that $N \geq M_{x_\infty}(\epsilon)$, so that the hypotheses of Proposition 3.4 are satisfied. Since $Qx_\infty = x_\infty$, we may write,

$$\mathcal{L}_N z_0 = N(x_\infty + \bar{o}(\epsilon)) + \bar{o}(c_1 M_{x_\infty}(\epsilon)) + \bar{o}(c_2)$$

Now, from Theorem 3.2(a) in [3] and the Banach Theorem,

$$\begin{aligned} ||v_N^* - z_0|| &\leq \frac{||\mathcal{L}_N z_0 - z_0||}{1-\lambda} \\ &= \frac{||N\bar{o}(\epsilon) + \bar{o}(c_1 M_{x_\infty}(\epsilon)) + \bar{o}(c_2)||}{1-\lambda}. \end{aligned}$$

If we now take $\epsilon = 1/N$ and substitute $z_0 = Nx_\infty$, the last result becomes,

$$||v_N^* - Nx_\infty|| \leq \frac{||\bar{o}(1) + \bar{o}(c_1 M_{x_\infty}(1/N)) + \bar{o}(c_2)||}{1-\lambda} \tag{4.2}$$

and this is satisfied for $N \geq M_{x_\infty}(1/N)$.

In light of Lemma 3.2, $M_{x_\infty}(1/N) = O(\log N)$. Hence, $N \geq M_{x_\infty}(1/N)$ and (4.2) holds for N sufficiently large. Moreover, the right hand side of (4.2) is $O(\log N)$. This yields (4.1) concluding the proof. □

The following Theorem documents some properties of x_∞.

Theorem 4.2 (Properties of x_∞) *Assume (C4). Then,*

(a) $||x_\infty||_{\text{sp}} = 0 \Leftrightarrow ||g^*||_{\text{sp}} = 0$

(b) x_∞ *is constant on the Bather classes.*

Proof.

(a) Substitute $x = c\mathbf{1}$ as a candidate solution into the equation $x = Qx$. Then the definition of Q implies

$$c\mathbf{1} = g^* + \lambda c\mathbf{1}.$$

A scalar c will render a solution in this vector equation iff g^* is state-independent.

(b) Fix any two states s_1, s_2 which are contained in the same Bather class. For any vector y, $\hat{U}^\infty y \in W$ and so is state-independent on each Bather class. In particular, letting $y = U^{a^\sigma} x_\infty$, it follows that $\hat{U}^\infty U^{a^\sigma} x_\infty(s_1)$ and $\hat{U}^\infty U^{a^\sigma} x_\infty(s_2)$ are maximized over A^σ by a common a_0^σ. Therefore

$$\begin{aligned} x_\infty(s_1) - x_\infty(s_2) &= (g^*(s_1) - g^*(s_2)) + \lambda(\hat{U}^\infty U^{a_0^\sigma} x_\infty(s_1) - \hat{U}^\infty U^{a_0^\sigma} x_\infty(s_2)) \\ &= 0 \end{aligned}$$

since both g^* and $\hat{U}^\infty U^{a_0^\sigma} x_\infty$ are state-independent on the Bather classes. □

5 Scaled ϵ-Optimality of Initially Stationary Policies

The following Theorem is the main result of this section. It establishes that $N\epsilon$-optimal i.s.p.'s exist whenever (C4) and (C5) hold. Moreover, the initial decision rule can be any γ satisfying (2.8) and so can be derived directly from Ψ.

Theorem 5.1 (Existence of scaled ϵ-optimal i.s.p.'s) *Assume (C4) and (C5) hold. Fix $\epsilon > 0$. Let $\eta_\epsilon = M_{x_\infty}((1-\lambda)\epsilon/4)$ and let γ be any decision rule satisfying (2.8). Then for all N sufficiently large, a uniform $N\epsilon$-optimal simple i.s.p. exists with planning horizon η_ϵ and initial decision rule γ.*

Proof.

Let $N > \eta_\epsilon$. For brevity, let $\epsilon' = (1-\lambda)\epsilon/4$. Fix $s \in S$ and an associated $a^\sigma \in A^\sigma$ such that

$$x_\infty(s) = g^*(s) + \lambda \hat{U}^\infty U^{a^\sigma} x_\infty(s). \tag{5.1}$$

Now, let $d \in D$ be a decision rule satisfying $U^{a^\sigma}\lambda x_\infty = U_d^{a^\sigma}\lambda x_\infty$. Let π denote a sequence of N decision rules whose first $N-\eta_\epsilon$ terms are γ and whose remaining terms satisfy $U_\pi^k U_d^{a^\sigma}\lambda x_\infty = U^k U^{a^\sigma}\lambda x_\infty$ for $0 \le k \le \eta_\epsilon$.

By Theorem 4.1,

$$\begin{aligned} L^{a^\sigma} v_N^* &= \max_{d\in D}\{r_d^{a^\sigma} + \lambda P_d^{a^\sigma}(Nx_\infty + O(\log N))\} \\ &= N\max_{d\in D}\{P_d^{a^\sigma}\lambda x_\infty\} + O(\log N) \\ &= NU^{a^\sigma}\lambda x_\infty + O(\log N). \end{aligned} \tag{5.2}$$

Similarly, by the definition of d,

$$L_d^{a^\sigma} v_N^* = NU^{a^\sigma}\lambda x_\infty + O(\log N). \tag{5.3}$$

From (5.2),

$$L^k L^{a^\sigma} v_N^* = L^k(NU^{a^\sigma}\lambda x_\infty) + O(\log N) \tag{5.4}$$

Applying Theorem (4.1) yet again gives, for $0 \le k \le \eta_\epsilon$ and N sufficiently large,

$$\begin{aligned} ||L^k L^{a^\sigma} v_N^* - L_\pi^k L_d^{a^\sigma} v_N^*|| &\le N||U^k U^{a^\sigma}\lambda x_\infty - U_\pi^k U_d^{a^\sigma}\lambda x_\infty|| \\ &\quad + 2\bar{r}(\eta_\epsilon + 1) + O(\log N) \\ &= N||U^k U^{a^\sigma}\lambda x_\infty - U_\pi^k U_d^{a^\sigma}\lambda x_\infty|| + O(\log N) \\ &= 0 + (1-\lambda)N\epsilon \end{aligned} \tag{5.5}$$

We have changed the the L and L^{a^σ} operators in the first inequality into U and U^{a^σ} operators at the expense of the term $2\bar{r}(\eta_\epsilon + 1)$. This term bounds the rewards obtainable in $k \le \eta_\epsilon + 1$ fastscale epochs. The third equality follows from the definition of π.

When $k > \eta_\epsilon$, Proposition 3.3 applies to the first term on the right hand side of (5.4) which then becomes,

$$\begin{aligned} L^k L^{a^\sigma} v_N^* &= kg^* + N(\hat{U}^\infty U^{a^\sigma}\lambda x_\infty + \bar{o}(\epsilon')) + \bar{o}(c_1\eta_\epsilon) + \bar{o}(c_2) + O(\log N) \\ &= kg^* + N\hat{U}^\infty U^{a^\sigma}\lambda x_\infty + N\bar{o}(\epsilon') + O(\log N). \end{aligned} \tag{5.6}$$

Applying $L_\pi^{\eta_\epsilon}$ to both sides of (5.3), we have

$$\begin{aligned} L_\pi^{\eta_\epsilon} L_d^{a^\sigma} v_N^* &= \bar{r}\bar{o}(\eta_\epsilon) + U_\pi^{\eta_\epsilon}(NU^{a^\sigma}\lambda x_\infty) + O(\log N) \\ &= \eta_\epsilon g^* + N\hat{U}^\infty U^{a^\sigma}\lambda x_\infty + N\bar{o}(\epsilon') + O(\log N). \end{aligned} \tag{5.7}$$

Reinvoking (3.8) with $\delta = \gamma$ and incorporating (5.7),

$$\begin{aligned} L_\pi^k L_d^{a^\sigma} v_N^* &= L_\gamma^{k-\eta_\epsilon} L^{\eta_\epsilon} L^{a^\sigma} v_N^* \\ &\geq (k-\eta_\epsilon)g^* + h_\gamma + P_\gamma^{k-\eta_\epsilon}(\eta_\epsilon g^* + N\hat{U}^\infty U^{a^\sigma}\lambda x_\infty) \\ &\quad + N\bar{o}(\epsilon') + O(\log N). \end{aligned}$$

Noting (C5), this becomes

$$\begin{aligned} L_\pi^k L_d^{a^\sigma} v_N^* &\geq (k-\eta_\epsilon)g^* + h_\gamma + \eta_\epsilon g^* + N\hat{U}^\infty U^{a^\sigma}\lambda x_\infty + N\bar{o}(\epsilon') + O(\log N) \\ &= kg^* + N\hat{U}^\infty U^{a^\sigma}\lambda x_\infty + N\bar{o}(\epsilon') + O(\log N) \end{aligned}$$

Subtracting this from (5.6) yields

$$L^k L^{a^\sigma} v_N^* - L_\pi^k L_d^{a^\sigma} v_N^* \leq N\bar{o}((1-\lambda)\epsilon/2) + O(\log N) \tag{5.8}$$

for $\eta_\epsilon < k < N$.

For the case $k = N$, we get a similar result by noting from (5.1) and Theorem 4.1 that

$$v_N^*(s) = N(g^*(s) + \lambda\hat{U}^\infty U^{a^\sigma} x_\infty(s)) + O(\log N)$$

and by subtracting (5.6) once again. This produces,

$$v_N^*(s) - L_\pi^N L_d^{a^\sigma} v_N^*(s) \leq N(1-\lambda)\epsilon/2 + O(\log N). \tag{5.9}$$

Taking (5.5), (5.8), and (5.9) together, it is clear that for N sufficiently large we will have, after taking norms

$$||L^k L^{a^\sigma} v_N^* - L_\pi^k L_d^{a^\sigma} v_N^*|| \leq (1-\lambda)N\epsilon$$

for all $0 \leq k \leq N-1$ and likewise that

$$v_N^*(s) - L_\pi^N L_d^{a^\sigma} v_N^*(s) \leq (1-\lambda)N\epsilon.$$

By Lemma 3.4 in [3], an i.s.p. which, for each initial states $s \in S$, selects a sequence $\{a^\sigma, \pi, d\}$ in this manner is uniform $N\epsilon$-optimal. Moreover, the structure of such an i.s.p. is as described in the statement of the Theorem. □

By letting $\eta_\epsilon = M_{x_\infty}(1/N)$, the preceding proof can be easily modified to prove the following Theorem.

Theorem 5.2 (Log order deviation from optimality) *Assume (C4) and (C5) hold. Let* $\eta(N)$ *be the diminishing function of order* $\log N$ *given by* $\eta(N) = M_{x_\infty}(1/N)$ *and let* γ *be any decision rule satisfying (2.8). Then a simple i.s.p. with planning horizon* $\eta(N)$ *and initial decision rule* γ *exists which is uniform* $O(\log N)$*-optimal.*

Hence, if we accept a diminishing planning horizon $\eta(N)$ as opposed to a bounded one η_ϵ, we can achieve approximate optimality up to a deviation term which is of order $\log N$ rather than a linear term $N\epsilon$.

6 ϵ-Optimality

In this section, we establish conditions for the existence of uniform ϵ-optimal i.s.p.'s. Our approach is the same as in [3], Section 5, namely to show that the sequences $\{L^k L^{a^\sigma} v_N^* - kg^*,\, a^\sigma \in A^\sigma\}$ approximately converge in $\eta < N$ steps. This leads to a proof that an ϵ-optimal i.s.p. exists with planning horizon η. From Example 8.3 in [3], we know that this η will generally not be bounded as a function of N. However, we may still hope for it to be a diminishing function. If so, we can show that ϵ-optimal i.s.p's exist with diminishing planning horizons.

The first obstacle in establishing such convergence is that, when g^* is state-dependent, the operators L and T are generally not equal. In reference [7], the authors describe 3 phases in which the general sequence $L^n x - ng^*$ converges. The first phase ends after a number of steps $n_0(x)$ when the L operator reduces to T. However, Example 1 in [7] shows that $n_0(x)$ can be linearly related to x. In our case, the initial reward vector x is $L^{a^\sigma} v_N^*$ which, due to Theorem 8.1(b) in [3], grows linearly in N. Hence, a sequence $L^k L^{a^\sigma} v_N^* - kg^*$ may not even complete the first phase of convergence within N steps, let alone converge in a diminishing number of steps.

Theorem 6.1 addresses this first obstacle. Part (a) establishes that, when (C4) and (C5) hold and N is sufficiently large, the L operator in $\{L^k L^{a^\sigma} v_N^* - kg^*,\, a^\sigma \in A^\sigma\}$ reduces not merely to T, but to $T_{(a^\sigma)}$ in a diminishing number of steps, $\eta(N)$.

Part(b) implies that, from that point onward, the sequences $\{L^k L^{a^\sigma} v_N^* - kg^*,\, a^\sigma \in A^\sigma\}$ evolve like sequences $\{T^k_{(a^\sigma)} O(\log N) - kg^*,\, a^\sigma \in A^\sigma\}$. The second obstacle is then to establish that these alternative sequences converge within ϵ in a diminishing number of steps $\eta_\epsilon(N)$. This is problematic because all that is known about the terminal rewards of this sequence is that they are $O(\log N)$ (and so potentially unbounded). Since $T_{(a^\sigma)}$ is an analogue of T, Theorem 4.2 in [7] indicates that that the convergence of $T^k_{(a^\sigma)} x - kg^*$ is geometric for each fixed terminal reward vector x. However, the geometric rate of convergence may not be uniformly bounded over an unbounded set of terminal rewards. This is true even in the state-dependent gain case, as Example 3 in [7] shows. Hence, we can anticipate nothing about the convergence time of the sequences $\{T^k_{(a^\sigma)} O(\log N) - kg^*,\, a^\sigma \in A^\sigma\}$.

The second obstacle was overcome in [2], Theorem B.4, which gave conditions such that the convergence time of T-operator value iteration is of at most the order of the terminal rewards. This allows us to complete the proof of the existence of uniform ϵ-optimal i.s.p.'s. The result is formerly presented in Theorem 6.2.

Theorem 6.1 (Reduction of L to $T_{(a^\sigma)}$) *Assume that (C4) and (C5) hold and let $\eta(N)$ be a diminishing function of order greater than $\log N$. Then if N is sufficiently large,*

(a) $\operatorname*{argmax}_{d\in D}\{r_d + P_d L^k L^{\eta(N)} L^{a^\sigma} v_N^*\} = \operatorname*{argmax}_{d\in D^{a^\sigma}}\{r_d + P_d L^k L^{\eta(N)} L^{a^\sigma} v_N^*\}$

(b) $L^k L^{\eta(N)} L^{a^\sigma} v_N^* = T^k_{(a^\sigma)} O(\log N) + \eta(N) g^* + N(\hat{U}^\infty U^{a^\sigma} \lambda x_\infty)$

for all $k = 0, 1, \ldots$ and $a^\sigma \in A^\sigma$.

Proof.

Fix any $a^\sigma \in A^\sigma$. By Theorem 4.1,

$$\begin{aligned} L^{a^\sigma} v_N^* &= \max_{d\in D}\{r_d^{a^\sigma} + \lambda P_d^{a^\sigma}(N x_\infty + O(\log N))\} \\ &= N \max_{d\in D}\{P_d^{a^\sigma} \lambda x_\infty\} + O(\log N) \\ &= N U^{a^\sigma} \lambda x_\infty + O(\log N) \end{aligned}$$

Hence, by the nonexpansive property of the L operator,

$$L^k L^{\eta(N)} L^{a^\sigma} v_N^* = L^{k+\eta(N)}(N U^{a^\sigma} \lambda x_\infty) + O(\log N).$$

By Lemma 3.2, $M_{x_\infty}(1/N) = O(\log N)$, while $\eta(N)$ is of order greater than $\log N$. We know, therefore that for N sufficiently large, $\eta(N) \geq M_{x_\infty}(1/N)$. In this case, Proposition 3.3 applies to the first term on the right hand side of the last equation with $G = G_{x_\infty}$. Hence, for all $k \geq 0$,

$$\begin{aligned} L^k L^{\eta(N)} L^{a^\sigma} v_N^* &= (k + \eta(N)) g^* + N(\hat{U}^\infty U^{a^\sigma} \lambda x_\infty + \bar{o}(1/N)) \\ &\quad + \bar{o}(c_1 M_{x_\infty}(1/N)) + \bar{o}(c_2) + O(\log N) \\ &= (k + \eta(N)) g^* + N(\hat{U}^\infty U^{a^\sigma} \lambda x_\infty) + O(\log N), \end{aligned} \tag{6.1}$$

from which we immediately get

$$\begin{aligned} &\operatorname*{argmax}_{d\in D}\{r_d + P_d L^k L^{\eta(N)} L^{a^\sigma} v_N^*\} \\ &= \operatorname*{argmax}_{d\in D}\{P_d[(k + \eta(N)) g^* + N(\hat{U}^\infty U^{a^\sigma} \lambda x_\infty)] + O(\log N)\}. \end{aligned} \tag{6.2}$$

Because arg max is unaffected by dividing its argument by a positive constant, we may modify the right hand side of (6.2) by dividing its argument by $\eta(N)$,

$$\begin{aligned} &\operatorname*{argmax}_{d\in D}\{r_d + P_d L^k L^{\eta(N)} L^{a^\sigma} v_N^*\} \\ &= \operatorname*{argmax}_{d\in D}\{P_d w(k,N) + O(\log N)/\eta(N)\} \end{aligned} \tag{6.3}$$

where we have introduced,

$$w(k,N) \triangleq \left(\frac{k+\eta(N)}{\eta(N)}\right) g^* + \left(\frac{N}{\eta(N)}\right) \hat{U}^\infty U^{a^\sigma} \lambda x_\infty.$$

Since (C5) holds, we know that W is convex (see Section 2.4). Noting that both g^* and $\hat{U}^\infty U^{a^\sigma}\lambda x_\infty$ are in W, it therefore follows from Lemma 3.1 (a) and (b) that $w(k,N) \in W$ and that $K(w(k,N)) = D^{a^\sigma}$ for all $k \geq 0$ and N. Moreover, since $\eta(N)$ is of order less than N, we may enlarge N so that $\left(\frac{N}{\eta(N)}\right) \geq 1$. In this case, Lemma 3.1(c) applies and

$$\Delta w(k,N) \geq \min(\Delta g^*, \Delta \hat{U}^\infty U^{a^\sigma} \lambda x_\infty)$$

for all $k \geq 0$ and N sufficiently large.

Therefore, by making yet another enlargement of N, (6.3) becomes

$$\begin{aligned} &\operatorname*{argmax}_{d\in D}\{r_d + P_d L^k L^{\eta(N)} L^{a^\sigma} v_N^*\} \\ &= \operatorname*{argmax}_{d\in D}\{P_d w(k,N) + \frac{1}{2}\bar{o}(\min(\Delta g^*, \Delta \hat{U}^\infty U^{a^\sigma} \lambda x_\infty))\} \\ &= \operatorname*{argmax}_{d\in D}\{P_d w(k,N) + \frac{1}{2}\bar{o}(\Delta w(k,N))\} \\ &= \operatorname*{argmax}_{d\in D^{a^\sigma}}\{P_d w(k,N) + \frac{1}{2}\bar{o}(\Delta w(k,N))\} \end{aligned} \tag{6.4}$$

where the last equality follows from (2.7).

Therefore $\operatorname*{argmax}_{d\in D}\{r_d + P_d L^k L^{\eta(N)} L^{a^\sigma} v_N^*\} \subseteq D^{a^\sigma}$ and, hence, part (a) follows for a fixed $a^\sigma \in A^\sigma$.

To prove part (b), observe that, by induction, part (a) implies

$$L^k L^{\eta(N)} L^{a^\sigma} v_N^* = T^k_{(a^\sigma)} L^{\eta(N)} L^{a^\sigma} v_N^* \tag{6.5}$$

for all $k \geq 0$ and N sufficiently large.

Letting $k = 0$ in (6.1) and noting that $K(\eta(N)g^* + N\hat{U}^\infty U^{a^\sigma}\lambda x_\infty) = D^{a^\sigma}$, we have

$$\begin{aligned} L^k L^{\eta(N)} L^{a^\sigma} v_N^* &= T^k_{(a^\sigma)}(O(\log N) + \eta(N) g^* + N\hat{U}^\infty U^{a^\sigma}\lambda x_\infty) \\ &= T^k_{(a^\sigma)} O(\log N) + \eta(N) g^* + N(\hat{U}^\infty U^{a^\sigma}\lambda x_\infty) \end{aligned}$$

for all $k \geq 0$ and N sufficiently large. Thus, part (b) is proved for fixed $a^\sigma \in A^\sigma$.

We have established that both (a) and (b) hold for all $k = 0, 1, 2, \ldots$ and N sufficiently large for an arbitrary fixed $a^\sigma \in A^\sigma$. The same conclusions follow for all $a^\sigma \in A^\sigma$ simultaneously since A^σ is finite. □

Theorem 6.2 (Existence of ϵ-optimal i.s.p.'s) *Assume (H1), (H2), and (C3) hold. Then for every fixed $\epsilon > 0$, there is a diminishing function $\eta_\epsilon(N)$ such that for all N sufficiently large, a uniform ϵ-optimal i.s.p. exists with planning horizon $\eta_\epsilon(N)$.*

Proof.

Hypothesis (H2) and (C3) together imply (C4). This is because each Bather class then contains a subchain of some aperiodic optimal decision rule. Because each $\hat{R}(\alpha)$ is communicating, a decision rule which randomizes over all actions not permitting a transition out of the Bather classes therefore has aperiodic chains $\hat{R}(\alpha), \alpha = 1, \ldots, \alpha^*$. Furthermore, (H1) and (H2) imply (C5).

Consequently Theorem 6.1 applies. As discussed previously, the question of whether the sequence $L^k L^{a^\sigma} v_N^*$ converges then reduces to the question of whether

$$T^k_{(a^\sigma)}(O(\log N)).$$

converges, for all $a^\sigma \in A^\sigma$.

In particular, if we can show that, for each fixed a^σ and ϵ, these sequences converge within a number of steps k which is $O(\log N)$. Then the total convergence time of $L^k L^{a^\sigma} v_N^*$ to within ϵ will be given by some $\eta_\epsilon(N)$ satisfying

$$\eta_\epsilon(N) = \eta(N) + O(\log N).$$

Here $\eta(N)$ refers to the diminishing function of order greater than $\log N$ described by Theorem 6.1. This sum is a diminishing function for each fixed ϵ. The Theorem then follows from arguments similar to those used to prove Theorem 5.1 in [3].

To show this, however, it is sufficient to demonstrate that assumptions (C3) and (A6) (see Appendix B in [2]) hold in each Ψ^{a^σ}. For then, by Theorem B.4 in [2], the convergence time of $T^k_{(a^\sigma)}(O(\log N))$ for each fixed ϵ is on the order of the terminal rewards, which is $\log N$. In the remainder of the proof, we argue that this is the case.

Fix $a^\sigma \in A^\sigma$. Note that when (H1) holds, any $d_0 \in D$ for which the Bather classes are closed will be an element of $K(w)$ for any $w \in W$. This is because w is constant on all the Bather classes, so naturally $P_{d_0} w = w$. In particular, any such d_0 will be in $K(g^*) \cap K(\hat{U}^\infty U^{a^\sigma} x_\infty) = D^{a^\sigma}$. However, decision rules which close the Bather classes are precisely the ones necessary to generate the different possible recurrent chains in Ψ. Clearly then, Ψ^{a^σ} has the same system of Bather classes and Schweitzer-Federgruen classes as Ψ. Therefore, (C3), (H1), and (H2) and consequently (C5) hold in Ψ^{a^σ}. Finally, as we argued in Section 2.5, E and D for the restricted MDP Ψ^{a^σ} are the same set, D^{a^σ}. Hence, in Ψ^{a^σ}, (C5) and (A6) are equivalent.

Since a^σ was arbitrary, we have shown that (C3) and (A6) hold in each Ψ^{a^σ}, completing the proof. □

Remark 6.3 If the hypotheses (H1) and (H2) are substituted with (H3), the Theorem still holds. The only difference in the proof is that the argument for why (A6) holds in each Ψ^{a^σ} is simpler. As explained in Section 2.3, all $w \in W$ are state-independent on the globally closed blocks of states under (H3). Hence, $D = K(w)$ for all $w \in W$. In particular, $D = E = D^{a^\sigma}$ for all $a^\sigma \in A^\sigma$, implying that (A6) holds for each Ψ^{a^σ}.

Theorem 6.2 can therefore be applied to all alternative projects models for which (H3) is satisfied.

7 Counter-examples

The results developed in previous sections require conditions (C4) and (C5). Similar to (C3) in [3], (C4) is a technical condition which excluded periodicity phenomena from our analysis. Generalizations which take periodicity into account are possible, although trite.

When (C5) is relaxed however, we find that PSMDPs may exhibit very different behavior than that described by the results in previous sections. One example where (C5) does not hold was presented in [3], namely Example 8.4. Here Equation (2.8) does not hold for $w \in W$ of the form $w = [\begin{array}{ccc} a & b & c \end{array}]^T$ with $a, c > b$. In that example, it was found that the convergence time of sequences $\{L^k L^{a^\sigma} v_N^* - kg^*,\, a^\sigma \in A^\sigma\}$ within arbitrary $\epsilon > 0$ may or may not be less than N. This depended jointly on the stationary and non-stationary data. When the convergence time was greater than N, the only cyclo-stationary policy which was $N\epsilon$-optimal for arbitrary ϵ was an i.s.p. with a non-average optimal initial decision rule.

In the following example, we find that

(i) Contrary to Theorems 5.1, 5.2, and 6.2, even a (uniform) $N\epsilon$-optimal i.s.p. whose planning horizon is diminishing does not exist.

(ii) Contrary to Theorem 6.1, the sequence $L^k L^{a^\sigma} v_N^* - kg^*$ does not converge, nor reach its second phase of convergence in a diminishing number of steps.

(iii) Despite (i), a (non-uniform) optimal i.s.p. does exist with a diminishing planning horizon, but is not identified by backward induction.

Example 7.1 Consider a DSM with

$$S = \{1,2,3\},\ A_1 = \{1\},\ A_2 = \{1,2\},\ A_3 = \{1\},\ \lambda = .75$$

and stationary data

$$\begin{array}{cccccc} s & a & r(s,a) & p(1|s,a) & p(2|s,a) & p(3|s,a) \\ \begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \end{bmatrix} & \begin{bmatrix} 1 \\ 1 \\ 2 \\ 1 \end{bmatrix} & \begin{bmatrix} 2 \\ 0.5 \\ -0.25 \\ -1 \end{bmatrix} & \multicolumn{3}{c}{\begin{bmatrix} 1 & 0 & 0 \\ 0.5-\mu & 2\mu & 0.5-\mu \\ 0.25 & 0 & 0.75 \\ 0 & 0 & 1 \end{bmatrix}} \end{array}$$

where $0 \le \mu < 0.5$. In addition, $A^\sigma = \{a^\sigma\}$ and

$$r_d^{a^\sigma} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},\quad P_d^{a^\sigma} = \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} & 0 \end{bmatrix}$$

for all $d \in D$. For simplicity, we shall take the parameter μ equal to zero. A more tedious development is possible for small, positive μ with the same conclusions.

We denote the decision rule which chooses action 1 in state 2 as δ and the decision rule which chooses action 2 as ζ. Both decision rules are aperiodic, so (C3) and (C4) hold. It is easily verified that δ is the only optimal decision rule in D and that $g^* = [\ 2\quad 0.5\quad -1\ ]^T$. Moreover, δ does not satisfy Equation (2.8) for all $w \in W$ (for instance $w = [\ 0\quad 0.75\quad 1\ ]^T$), implying that (C5) does not hold.

We will show by direct substitution that the vector $N[\ 2\quad 1\quad 0\ ]^T$ solves the optimality equation and, hence, that $v_N^* = N[\ 2\quad 1\quad 0\ ]^T$. We let

$$y_0 = L^{a^\sigma}([\ 2\quad 1\quad 0\ ]^T) = N[\ 0\quad 0.75\quad 1\ ]^T$$

and examine the sequence $L^k y_0,\ k = 0,1,2,...,N$.

By induction, it may be verified that, for $k < N/3+1$, the only improving decision rule in this sequence is ζ and

$$L^k y_0 = N \begin{pmatrix} 0 \\ 0.75 \\ 1 \end{pmatrix} + k \begin{pmatrix} 2 \\ -0.25 \\ -1 \end{pmatrix} \tag{7.1}$$

for $k \le N/3+1$.

Using this result, we can use induction once again to show that, for $k > N/3+1$, the only improving decision rule is δ and

$$L^k y_0 = L_\delta^k y_0 = N \begin{pmatrix} 0 \\ 0.5 \\ 1 \end{pmatrix} + k \begin{pmatrix} 2 \\ 0.5 \\ -1 \end{pmatrix}. \tag{7.2}$$

In the particular case $k = N$,

$$L^N y_0 = N[\ 2 \quad 1 \quad 0\ ]^T = \mathcal{L}_N(N[\ 2 \quad 1 \quad 0\ ]^T)$$

proving that $N[\ 2 \quad 1 \quad 0\ ]^T$ solves the optimality equation.

As we noted, δ only becomes an improving decision rule after $N/3+1$ value iteration steps, which is non-diminishing. Furthermore, it is easily verified that $E = \{\delta\}$. Hence, L reduces to T only after a non-diminishing number of steps.

Now let

$$e(k, y_0) \stackrel{\triangle}{=} L^k y_0 - kg^* - N[\ 0 \quad 0.5 \quad 1\ ]^T.$$

We deduce from (7.1) that

$$e(k, y_0) = N \begin{pmatrix} 0 \\ 0.25 \\ 0 \end{pmatrix} + k \begin{pmatrix} 0 \\ -0.75 \\ 0 \end{pmatrix}$$

for $k \leq N/3+1$. From (7.2) we deduce that $e(k, y_0) = 0$ for $k > N/3+1$. This directly indicates that

$$\hat{L}^\infty y_0 = \hat{L}^\infty L^{a^\sigma} v_N^* = N[\ 0 \quad 0.5 \quad 1\ ]^T$$

and that the sequence $L^k L^{a^\sigma} v_N^*$ converges within an arbitrary error only after a non-diminishing number of steps $k = N/3+1$.

We now demonstrate, by contradiction, that it is impossible, for all N sufficiently large, to achieve (uniform) $N\epsilon$-optimality for arbitrary ϵ via an i.s.p. with a diminishing planning horizon. For suppose it were possible with diminishing planning horizon function, $\eta(N)$. Then for any ϵ, we may extract a subsequence N_j and an associated sequence of $N_j\epsilon$-optimal i.s.p.'s π_j, each having planning horizon $\eta(N_j)$ and the same initial decision rule.

Take $\epsilon = 0.1$. Suppose first, that the common initial decision rule for the sequence is δ. Now, fix j sufficiently large so that $\eta(N_j) < N_j/3+1$. This condition can be satisfied since $\eta(N)$ is diminishing. Using (7.1), we find, by direct substitution, that

$$\begin{aligned} & L^{\eta(N_j)+1} L^{a^\sigma} v_N^* - L_\delta L^{\eta(N_j)} L^{a^\sigma} v_N^* = \\ = \quad & N_j \begin{pmatrix} 0 \\ 0.25 \\ 0 \end{pmatrix} + (\eta(N_j)+1) \begin{pmatrix} 0 \\ -0.75 \\ 0 \end{pmatrix}. \end{aligned} \tag{7.3}$$

If we consider starting the process $\eta(N_j)+1$ steps before a renewal epoch, then since each π_j is uniform $N_j\epsilon$-optimal,

$$\begin{aligned} N_j\epsilon &\geq L^{\eta(N_j)+1}L^{a^\sigma}v_N^*(2) - L_{\pi_j}^{\eta(N_j)+1}L^{a^\sigma}v^{\pi_j}(2) \\ &\geq L^{\eta(N_j)+1}L^{a^\sigma}v_N^*(2) - L_\delta L_{\pi_j}^{\eta(N_j)}L^{a^\sigma}v_N^*(2) \\ &\geq L^{\eta(N_j)+1}L^{a^\sigma}v_N^*(2) - L_\delta L^{\eta(N_j)}L^{a^\sigma}v_N^*(2). \end{aligned}$$

The last two inequalities used the monotonicity of dynamic programming operators.

Combining this with (7.3) and substituting $\epsilon = 0.1$ yields

$$(N_j)0.1 \geq L^{\eta(N_j)+1}L^{a^\sigma}v_N^*(2) - L_\pi^{\eta(N_j)+1}L^{a^\sigma}v_N^*(2) \geq 0.25N_j - 0.75(\eta(N_j)+1)$$

Dividing through by N_j and letting j tend to infinity produces $0.1 \geq 0.25$ thereby establishing a contradiction.

Alternatively, if we suppose that the initial decision rule common to the sequence is ζ, then using a similar approach, we first write

$$v_N^* - L_\zeta^{N_j-\eta(N_j)}L^{\eta(N_j)}L^{a^\sigma}v_N^* = N[\ 0 \quad 0.5 \quad 0\ ]^T.$$

The fact that π_j is $N_j\epsilon$-optimal, implies

$$\begin{aligned} N_j\epsilon &\geq v_N^*(2) - L_\zeta^{N_j-\eta(N_j)}L^{\eta(N_j)}L^{a^\sigma}v_N^*(2) \\ &= 0.5N_j. \end{aligned}$$

In this case, the contradiction $0.1 \geq 0.5$ is reached.

Although we have proved that uniform optimal i.s.p.'s can not have diminishing planning horizons in this model, it is nevertheless true that the i.s.p. which uses δ at every fastscale epoch is conventionally optimal. This follows from the second equality in (7.2). When $k = N$,

$$L_\delta^N L^{a^\sigma}v_N^* = N[\ 2 \quad 1 \quad 0\ ]^T = v_N^*.$$

Hence, a conventionally optimal i.s.p. exists with the diminishing planning horizon $\eta = 0$.

8 Conclusions

We have analyzed PSMDP's whose underlying MDP have state-dependent optimal gain. The properties of i.s.p.'s are not as strong in this case as in the state-independent gain case. This is due to the unboundedness of the optimal value together with the irregular behavior of value iteration in this case. Examples have shown that, in general, optimal and ϵ-optimal i.s.p.'s may not have bounded planning horizons. When using i.s.p.'s, it is necessary to settle for scaled ϵ-optimality and/or diminishing planning horizons.

Despite its more intricate nature, much information about the optimal discounted reward and the optimality of i.s.p.'s was obtained in the state-dependent gain case. The results rested on conditions which are naturally satisfied, e.g., in application of the alternative projects management type where the projects may be individually modeled as weakly communicating MDPs. Theorem 4.1 gave us a decomposition of v_N^* into an explicit term Nx_∞ and a term of order $\log N$. Also, Theorem 5.1 guaranteed the existence of an $N\epsilon$-optimal simple i.s.p. Theorem 5.2 guaranteed a similar i.s.p. which is $O(\log N)$-optimal, but which has a diminishing planning horizon.

Proving the existence of ϵ-optimal (non-scaled) i.s.p.'s with diminishing planning horizons was by far the most challenging problem handled here. An important property required in the analysis is the convergence of relevant value iteration sequences in a diminishing number of backward induction steps. Established value iteration theory suggests that, in general, these sequences might not even complete their first phase of convergence in a number of steps which is less than N. Theorem 6.1, however, showed that, under (C4) and (C5), the sequences reach their second phase within a diminishing number of steps. Moreover, the terminal reward of the sequence is effectively $O(\log N)$. The remainder of the convergence problem was resolved through Theorem B.4 in [2]. The Theorem established conditions such that the convergence time of second-phase value iteration is comparable to the size of the terminal rewards.

The work of this paper and its prequel [3] identify conditions under which the structure of optimal policies is strongly influenced by the limiting behavior of the stationary data's value iteration operator. When these conditions are relaxed, examples have shown that this may not be the case. Phenomena such as i.s.p.'s with non-gain optimal initial decision rules may then be observed. The investigation of such phenomena remains open. It seems evident, however, that it will require tools beyond the theory of average optimality and the limiting behavior of non-discounted value iteration. Moreover, conditions on the non-stationary data may play a greater role than in our analysis.

Another future direction for analysis of PSMDPs is alternative optimality criteria. The $(N+1)$-step discounted optimality problem which we have formulated is analogous to a standard discounted MDP with stationary probabilities and rewards. With little difficulty, one can see how other analogue $(N+1)$-step criteria (e.g. average optimality, Blackwell optimality, etc...) may also be formulated for PSMDPs. For all of these criteria, optimality will also be achieved on the class of cyclo-stationary policies. Here again, we anticipate that more involved analysis will be required – including combined consideration of both the stationary and non-stationary chain structure – than for the present discounted problem.

References

[1] J. Bather (1973) Optimal decision procedures for finite Markov chains III. *Adv. Appl. Prob.* **5** 541-553.

J. Math. Anal. Appl. **65** 711-730.

[2] M. Jacobson *Asymptotic Properties of Two Timescale Markov Decision Processes.* M.Sc. Thesis, Technion - Israel Institute of Technology, October 1998. Available at: http://www.ee.technion.ac.il/~adam/PAPERS/MWJ-MSc.ps

[3] M. Jacobson, N. Shimkin, A. Shwartz (1999) *Piecewise Stationary Markov Decision Processes, I: Constant Gain.* Submitted for publication. Available at: http://www.ee.technion.ac.il/~adam/PAPERS/2TimesScale1.ps

[4] Puterman, Martin L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc. 1994.

[5] P.J. Schweitzer, and A. Federgruen (1978). The functional equations of undiscounted Markov renewal programming. *Math Opns Res.* **3** 308-21.

[6] P.J. Schweitzer and A. Federgruen (1977). The asymptotic behavior of undiscounted value iteration in Markov decision problems. *Math Opns Res.* **2** 360-82.

[7] P.J. Schweitzer and A. Federgruen (1979). Geometric convergence of value iteration in multichain Markov decision processes. *Adv Appl Prob.* **11** 188-217.

[8] P.J. Schweitzer (1984). A Value Iteration Scheme for Undiscounted Multichain Markov Renewal Programs. *Zeitschrift für Operations Research* **28** 143-52.

Notation List (Attached for the Reviewers' Convenience)

a	A fastscale action
a^σ	A slowscale action
A_s	The set of allowable fastscale actions in state s
A^σ	The set of available slowscale actions
d	A decision rule for choosing fastscale actions
D	The set of deterministic decision rules by which fastscale actions can be chosen; hence, also the set of decision rules in the underlying MDP.
D^*	The set of gain optimal decision rules in the underlying MDP
D^{a^σ}	The set of decision rules given by $E \cap K(\hat{U}^\infty U^{a^\sigma} x_\infty)$ for any slowscale action $a^\sigma \in A^\sigma$
E	The subset of decision rules $d \in D$ satisfying $P_d g^* = g^*$
g^*	Optimal gain vector of the underlying MDP
G_{x_∞}	The set of vectors defined as $G_{x_\infty} \stackrel{\triangle}{=} \{x \in \mathbb{R}^{\lvert S \rvert} \mid x = U^{a^\sigma} \lambda x_\infty,\ a^\sigma \in A^\sigma\}$
$K(w)$	For $w \in W$, the decision rules $d \in D$ satisfying $Uw = P_d w$
L	The single step dynamic programming operator for the stationary data; $Lx = \max_{d \in D}\{r_d + P_d x\}$
L_d	The restriction of the L operator to the decision rule $d \in D$
L_π^k	The k-th reward-to-go function for π, a finite sequence of decision rules in D whose length is at least k
$\hat{L}^\infty$	Operator defined via the limit $\hat{L}^\infty x = \lim_n L^n x - ng^*$
L^{a^σ}	Discounted dynamic programming operator for the non-stationary data and slowscale action $a^\sigma \in A^\sigma$; $L^{a^\sigma} x = \max_{d \in D}\{r_d^{a^\sigma} + \lambda P_d^{a^\sigma} x\}$
$\mathcal{L}_N$	The discounted dynamic programming operator for a PSMDP
$M_U(G, \epsilon)$	The number of steps required for the sequence $U^n x$ to converge to within ϵ for all $x \in G$ where G is a finite set of terminal rewards
$M_{x_\infty}(\epsilon)$	The function defined as $M_{x_\infty}(\epsilon) \stackrel{\triangle}{=} M_U(G_{x_\infty}, \epsilon)$
N	The number of stationary epochs in each renewal cycle of a PSMDP
$\bar{o}(c)$	An unspecified vector whose sup norm is bounded by c
Q	Operator defined via $Qx(s) = \max_{a^\sigma \in A^\sigma}\{g^* + \lambda \hat{U}^\infty U^{a^\sigma} x(s)\}$

	for vectors x and all $s \in S$
$\hat{R}$	The set of states which are recurrent for some decision rule in the underlying MDP (the Bather classes)
R^*	The set of states which are recurrent for some average optimal decision rule in the underlying MDP (the Schweitzer-Federgruen classes)
S	The state space
T	The restriction of L to decision rules in E
$\hat{T}^\infty$	Operator defined via the limit $\hat{T}^\infty x = \lim_n T^n x - ng^*$
$T_{(a^\sigma)}$	The restriction of L to decision rules in D^{a^σ}
U	A single step dynamic programming operator when the reward data for the underlying MDP is zero; $Ux = \max_{d \in D}\{P_d x\}$
$\hat{U}^\infty$	Operator defined via the limit $\hat{U}^\infty x = \lim_n U^n x$
U^{a^σ}	A version of U for the non-stationary data; $U^{a^\sigma} x = \max_{d \in D}\{P_d^{a^\sigma} x\}$
v	An element of V
v_N^*	The optimal value vector of a discounted PSMDP
V	The set of solutions to the second average optimality equation in the underlying MDP; $V = \{v \mid g^* + v = Tv\}$
w	An element of W
W	A version of V when the rewards in the underlying MDP are zero; $W = \{w \mid w = Uw\}$
x_∞	The fixed point of Q
γ	An average optimal decision rule in the underlying MDP satisfying $P_\gamma w = w$ (as in Assumption (C5))
Δw	For $w \in W$, the closest that a decision rule $d \in D$ can bring $P_d w$ to Uw without actually attaining the maximum
$\eta(N)$	An N-dependent planning horizon; a diminishing function of N
λ	The discount factor of the discounted PSMDP
Ψ	The underlying MDP
Ψ^{a^σ}	A version of Ψ obtained by restricting D to D^{a^σ}
$\mathbf{1}$	A column vector whose every element is one