Piecewise Stationary Markov Decision Processes, I: Constant 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

We consider a class of non-stationary dynamic decision-making models to which we refer as Piecewise Stationary Markov Decision Processes (PSMDPs). In these models, the decision making horizon can be partitioned into intervals, called *renewal cycles*, of $N+1$ epochs. The transition law and reward function are identical over the first N epochs of each renewal cycle, but distinct at the final epoch. The motivation for these models is in applications where decisions of different nature are taken at different time scales, i.e. many "low-level" decisions are made between "high-level" ones.

Our aim is to characterize solutions of the discounted reward optimality problem for large values of N, with the effective discount rate over an entire renewal cycle held fixed. In this model, *initially stationary* policies are natural candidates for optimal policies. Similar to turnpike policies, an initially stationary policy uses a fixed decision rule for some large number of epochs in each renewal cycle, followed by a relatively short planning horizon of time-varying decision rules.

Our analysis relates the PSMDP to an average reward MDP defined by the stationary part of the system. We focus here on the constant gain case, where this MDP's optimal average reward is independent of the initial state. We find that the optimal value of the PSMDP can be fully characterized. It is shown that initially stationary policies are ϵ-optimal under weak conditions and require a planning horizon whose length is bounded in N. We further identify conditions under which these policies are precisely optimal. The non-constant gain case is briefly considered here via examples, and further analyzed in the companion paper [7].

1 Introduction

The work presented here concerns sequential decision-making problems in which decisions of different nature are taken at different time scales. Such applications lead to decision process models similar to Markov Decision Processes (MDPs), but which possess a particular piecewise stationary structure. In these models, the decision horizon may be 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 epochs*), the rewards and transition functions are distinct. These distinct functions describe the effect of decisions made at a slower time scale. We refer to such types of decision processes as Piecewise Stationary Markov Decision Processes (PSMDPs).

In this paper, we define two sub-categories of PSMDPs which we respectively term Markovian Slowscale Models (MSMs) and Delayed Slowscale Models (DSMs). In an MSM, all transitions and rewards are described by standard, Markovian probability and reward functions. At non-stationary epochs these functions are different from those at stationary epochs. In a DSM, meanwhile, rewards and transitions at non-stationary epochs are determined not only by the current state and action, but also by an action (of a different type) which is taken at the start of the current renewal cycle. Hence, there is a delay of many epochs between the time when the action of the second type is taken and when it affects state transitions and rewards. This means that DSMs, unlike MSMs, have a history-dependent structure.

The DSM and MSM models are quite similar, but have conventions suiting different applications. MSMs are highly suited to applications where routine decision making is periodically interrupted by non-routine, higher-level decisions. An example is the management of multiple project operations, in which the decision maker periodically selects one of several projects to work on for a term of N epochs. In this case, the actions taken at non-stationary epochs correspond to selections of projects, while decisions at intervening epochs determine the evolution of the current project.

In such applications, the projects may be likened to the discrete component of a hybrid system. In hybrid systems (see, for example, [1, 2, 3, 4]), a discrete state component, indicating the mode of the system, makes infrequent transitions relative to a second, continuous state component. The continuous component evolves, between these transitions, according to state equations determined by the current mode.

The motivation for DSMs is in applications where the effect of certain types of actions on the physical variables in the system is slower to arrive than other types. If decision epochs represent the times where the "faster" actions are taken, then there is a lag of many epochs between the epochs when the "slower" actions are taken and when they take effect. As an example, one might consider a manufacturing plant which produces goods by the hour using some stock of manufacturing components. If we suppose that supply orders of the manufacturing components are made monthly *and* take a month to process, then the plant may be appropriately modeled as a DSM in which the fast actions are manufacturing decisions and the slow actions are order levels for new components.

For PSMDPs of these two types, we consider the problem of optimizing the total expected reward when rewards devalue by a discount factor λ at the beginning of each renewal cycle. Under this criterion, the renewal cycle length is regarded as the time scale on which the present value of rewards changes. Our aim is to understand how the solutions of this problem behave for large, finite values of N, i.e. when the interval of stationary epochs in each renewal cycle is long. We assume finite state and action spaces throughout.

As usual, we are interested both in the optimal value obtainable and the decision-making policies which attain it. We expect that for large N, a policy of a simplified form, which we call an *initially stationary policy* (i.s.p.), will result in near optimality. An i.s.p. is a policy which, in each renewal cycle, uses a fixed decision rule for some large number of epochs followed by a relatively short horizon of time-varying decision rules.

This structure is highly reminiscent of policies associated with turnpikes. The notion of turnpikes originates in the work of Shapiro [13], and was subsequently expanded by Hinderer and Hubner [5]. It is rooted in the intuition that, at the start of a stationary MDP with a large, finite horizon, it appears to the decision maker that the horizon before him is infinite. Initially, therefore, optimal decisions can be made using a time-invariant decision rule, as in an infinite-horizon version of the MDP. As the end of the horizon approaches, however, the problem looks more and more like a finite horizon problem and time-varying decision rules may be required.

This intuition was verified in [5, 13] for discounted MDPs possessing a unique optimal decision rule. The time-invariant decision rule is referred to as a "turnpike", while the interval of time-varying decision rules is referred to as the *planning horizon.* Analogous results hold for non-discounted MDPs as well. The same intuition leads us to conjecture that an initially stationary policy might be optimal in PSMDPs. For in these models, the decision maker faces a long stationary horizon at the start of each renewal cycle.

The analysis in this paper is carried out under certain conditions on Ψ, the average reward MDP associated with the stationary data. The basic requirement is that the optimal gain of Ψ is independent of the initial state. Analysis of the PSMDP starts by characterizing the form of the optimal value as a function of N. We establish (Theorem 4.1) that the optimal value is approximately the sum of two terms. The first term is proportional to N and the gain of Ψ, and does not depend on the non-stationary data. The second term is bounded independently of N and converges to a limiting vector as N increases. This structure is used to establish the existence of ϵ-optimal initially stationary policies with a finite planning horizon which depends only on ϵ but not on N. It is worth noting that these results hold under mild technical conditions on Ψ; in particular, they require no conditions on the rate of convergence of value iteration in Ψ in order to bound the planning horizon.

Moreover, under some additional technical conditions (which hold, for example, if Ψ is unichain), we establish that, even in the DSM model, ϵ-optimal i.s.p.'s have a simple Markovian structure. Namely, actions are chosen based on the current state and time alone. This is a noteworthy finding for DSMs since they have a history-dependent structure. Existence of precisely optimal i.s.p.'s is examined next, and established under the additional requirement that the set of average optimal decision rules which are maximizing in Ψ's optimality equation is a singleton.

The rest of this paper is organized as follows. Section 2 includes a model introduction and defines notation which we work with throughout. Section 3 describes the optimality criteria and the relevant optimality equations for PSMDPs. Since the structure of the DSM optimality problem is seen to be readily extendible to MSMs, our analysis in this and subsequent sections is restricted to the former, while the analogous results for MSMs are summarized in Section 7.

In Sections 4, 5, and 6, we treat DSMs with state-independent gain. Section 4 establishes the asymptotic form of the value function, while Sections 5 and 6 study ϵ-optimal and optimal i.s.p.'s respectively. Finally, in Section 8, a preliminary examination is made of PSMDPs for which Ψ has state-dependent gain. This is to develop a sense of the more complex behavior of such models as compared to the state-independent gain case. We first establish that the optimal value is bounded in N, with respect to the span norm, if and only if Ψ has state-independent optimal gain. Example 8.3 illustrates that ϵ-optimal i.s.p.'s may require unbounded planning horizons. Example 8.4 illustrates that ϵ-optimal i.s.p.'s may require initial decision rules which are not gain optimal in Ψ. A fuller analysis of the state-dependent gain case may be found in a sequel paper [7].

2 Model Introduction and Notations

2.1 Model Introduction

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 $t = 0, (N+1), 2(N+1), \ldots$ – *renewal epochs.* The first N epochs of each renewal cycle shall be called *stationary epochs.* Accordingly, the reward and transition probability functions governing

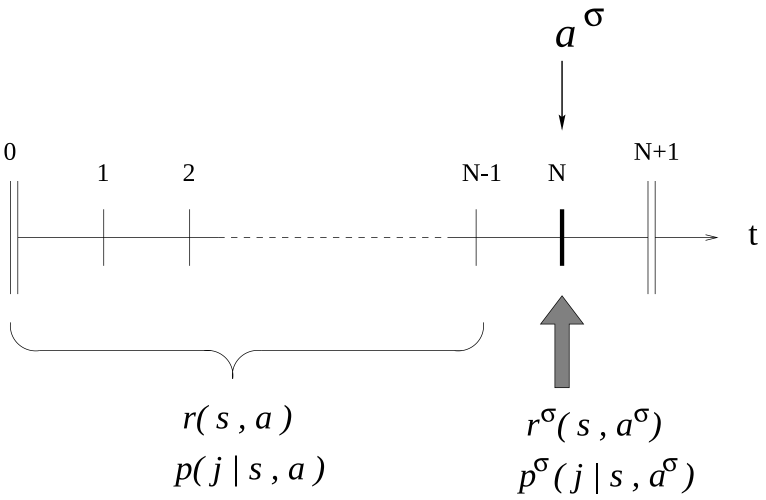

Figure 1: Time evolution of a Markovian Slowscale Model (MSM)

the stationary epochs shall be referred to as the *stationary data.* Conversely, we shall call the final epoch in each renewal cycle the *non-stationary epochs* and refer to the relevant reward and transition probability functions as the *non-stationary data.*

We furthermore define two sub-categories of PSMDPs, respectively called Markovian Slowscale Models (MSMs) and Delayed Slowscale Models (DSMs).

An MSM is specified by a state space S and, for each $s \in S$, spaces of allowable *fastscale* actions A_s and *slowscale* actions A_s^σ. Fastscale actions are selected at stationary epochs where rewards and transition probabilities are given by stationary data $r(s_t, a_t)$, $p(s_{t+1}|s_t, a_t)$. Slowscale actions are taken at non-stationary epochs where rewards and transitions are given by non-stationary data $r^\sigma(s_t, a_t^\sigma)$, $p^\sigma(s_{t+1}|s_t, a_t^\sigma)$. The MSM timing framework is depicted for one renewal cycle in Figure 1.

A DSM is specified by a state space S, a space of allowable fastscale actions A_s for each state $s \in S$, and a space A^σ of slowscale actions [1]. In this case, 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. As in MSMs, 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 $r(s_t, a_t, a^\sigma_{(N+1)\lfloor t/(N+1)\rfloor})$, $p(s_{t+1}|s_t, a_t, a^\sigma_{(N+1)\lfloor t/(N+1)\rfloor})$. The DSM timing framework is depicted for one renewal cycle in Figure 2.

For a PSMDP, we define the history, h_t, as the chronological sequence of states observed and actions taken ending with the state observed at epoch t. A randomized history-dependent policy

[1] As in MSMs, it is possible to let the set of allowable slowscale actions vary with s, but no interesting generalizations seem to follow from this.

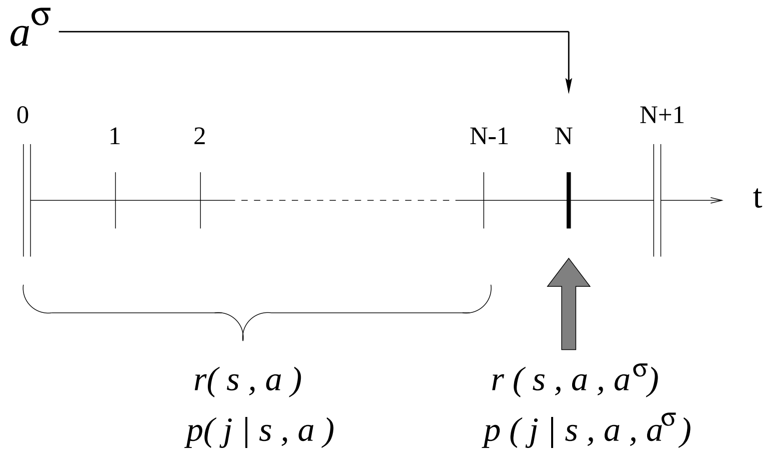

Figure 2: Time evolution of a Delayed Slowscale Model (DSM)

is a sequence of functions $q_{h_t}(\cdot)$ on the set of decisions possible at t. These functions indicate the probability of making a particular decision when the history is h_t.

Let $\mathbf{\Pi}_N^{HR}$ denote the set of all randomized history-dependent policies, for a certain N. Each $\boldsymbol{\pi} \in \mathbf{\Pi}_N^{HR}$ induces, for each initial state $s \in S$, a probability measure P_s^{π} – and a corresponding expectation operator E_s^{π} – on the space of possible histories h_t. The construction of this probability measure is standard and is based on the law of conditional probabilities.

We shall assume throughout that all state and action spaces are finite.

2.2 Norms and Dynamic Programming Notation

We shall represent functions on S by column vectors in $\mathbb{R}^{|S|}$. Likewise, we shall define various dynamic programming operators and think of them as mappings of vectors from $\mathbb{R}^{|S|}$ to $\mathbb{R}^{|S|}$.

Let us introduce the following notation for a given vector $y \in \mathbb{R}^{|S|}$.

$$\begin{aligned} ||y|| &\stackrel{\triangle}{=} \max_{s\in S} |y(s)| \\ ||y||_{\mathrm{sp}} &\stackrel{\triangle}{=} \max_{s\in S} y(s) - \min_{s\in S} y(s). \end{aligned}$$

The span semi-norm $||y||_{\mathrm{sp}}$ measures the maximal variation between the elements of y.

For a scalar $c \geq 0$, we use the notation $\bar{o}(c)$ to denote an unspecified vector in $\mathbb{R}^{|S|}$ whose sup-norm is at most c.

Let L denote the one-step dynamic programming operator for the stationary data, defined as

$$Lx(s) = \max_{a\in A_s}\{r(s,a) + \textstyle\sum_{j\in S} p(j|s,a)x(j)\}, \quad s \in S \ .$$

Some well known properties of the L operator are

(i) $L(x+c\mathbf{1}) = Lx + c\mathbf{1}$ where c is any scalar and $\mathbf{1}$ denotes a column vector whose every element is 1.

(ii) $||Lx - Ly|| \leq ||x-y||$, $||Lx - Ly||_{\text{sp}} \leq ||x-y||_{\text{sp}}$.

Property (ii) is sometimes referred to as the *non-expansive property* of the L operator. From this property, we deduce that $L(y+\bar{o}(c)) = Ly + \bar{o}(c)$ for any vector y.

For an MSM and a given $0 \leq \lambda < 1$, we define the discounted dynamic programming operator for the non-stationary data

$$L^{\sigma}x(s) = \max_{a^{\sigma} \in A_s^{\sigma}} \{r^{\sigma}(s,a^{\sigma}) + \lambda \sum_{j \in S} p^{\sigma}(j|s,a^{\sigma})x(j)\}, \quad s \in S \ .$$

Similarly, for a DSM, each $a^{\sigma} \in A^{\sigma}$ shall be associated with a discounted operator

$$L^{a^{\sigma}}x(s) = \max_{a \in A_s} \{r(s,a,a^{\sigma}) + \lambda \sum_{j \in S} p(j|s,a,a^{\sigma})x(j)\}, \quad s \in S$$

for the non-stationary data. Finally, we define the $(N+1)$-step DSM discounted operator $\mathcal{L}_N$,

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

It corresponds to backward induction over one renewal cycle.

2.3 Policies and Decision Rules

A *decision rule* is a mapping from states to actions, and can be thought of as a rule for selecting actions at some epoch based on the current state. Let D denote the space of decision rules mapping each $s \in S$ to some $a \in A_s$. In MSMs, D^{σ} shall denote the space of decision rules mapping $s \in S$ to $a^{\sigma} \in A_s^{\sigma}$.

For a decision rule $d \in D$, we define the restriction of L to d as

$$L_d x \stackrel{\triangle}{=} r_d + P_d x,$$

where r_d is a reward vector with components $r_d(s) = r(s,d(s))$ and P_d is a transition probability matrix with entries $P_d(j,s) = p(j|s,d(s))$. Restrictions of other operators are similarly defined.

We shall sometimes represent L using the vector form,

$$Lx = \max_{d \in D} \{r_d + P_d x\},$$

where the maximization over D is component-wise, and similarly for other dynamic programming operators. Note that $Lx \geq L_d x$ with equality for at least one d. A decision rule, d, attaining this maximum shall be called an x-improving or x-maximizing decision rule.

If $\pi = \{d_m, d_{m-1}, \dots, d_1\}$ represents a sequence of m decision rules, then for all $0 \le k \le m$,

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

In other words $L_\pi^k x$ denotes the k-th reward-to-go function of the m-horizon policy π when the terminal reward is x.

The maxima on the right hand side of (2.1) are obtained by s-dependent sequences of the form

$$\{a^\sigma, d_0, d_1, \dots, d_N\}(s), \quad s \in S \tag{2.2}$$

such that

$$\mathcal{L}_N x(s) = L_{d_0} L_{d_1} \cdots L_{d_{N-1}} L_{d_N}^{a^\sigma} x(s).$$

A set of sequences of the form (2.2) can be associated with a DSM policy which, at a given renewal epoch, *prescribes* the sequence $\{a^\sigma, d_0, d_1, \dots, d_N\}(s)$ if the current state is s. The element a^σ is the slowscale action chosen at the renewal epoch while the decision rule $d_i, i = 0, 1, \dots, N$ is the rule that will be used to select fastscale actions at the i-th epoch of the current renewal cycle. We call policies of this form cyclo-stationary policies because of their $(N+1)$-periodic form. The set of all cyclo-stationary policies shall be denoted $\mathbf{\Pi}_N^{\text{cyc}}$.

A sequence prescribed in some state by a cyclo-stationary policy, shall be abbreviated as $\{a^\sigma, \pi, d\}$ where $\pi = \{d_0, d_1, \dots, d_{N-1}\}$ are the decision rules used at the stationary epochs while d is the decision rule used at the non-stationary epoch.

A particular type of cyclo-stationary policy which will be of interest to us is an initially stationary policy (i.s.p.) which has the form

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

I.e., when a renewal cycle starts in state $s \in S$, the policy prescribes some fixed decision rule, δ (which may depend on s), for all but possibly the final η stationary epochs ($\eta < N$) of the renewal cycle. We refer to η as the i.s.p.'s *planning horizon* [2] (borrowing from turnpike terminology). Also, we refer to δ as the *initial decision rule* for the particular state s. If the initial decision rule is the same for all s, we call the i.s.p. *simple.*

2.4 The Underlying MDP

The stationary data in a PSMDP can be associated with an average reward MDP, denoted Ψ, with decision rule space D. We shall refer to Ψ as the *underlying MDP* of the PSMDP.

[2] We shall sometimes use the term planning horizon loosely to mean also the *interval* of η epochs in a renewal cycle where time-varying decision rules are used by an i.s.p.

The gain vector g_d of a decision rule $d \in D$ is defined as $g_d \triangleq P_d^* r_d$ where $P_d^* = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} P_d^k$ is the limiting matrix of P_d. Since the action and state spaces are finite, it is known that g^*, the optimal gain vector of Ψ, satisfies $g^* \geq g_d$ with equality for some decision rule d. We define $D^* \triangleq \{d \in D \mid g^* = g_d\}$ as the set of optimal decision rules in Ψ.

The optimal gain vector, g^*, is characterized by the optimality equations

$$\begin{aligned} g^* &= \max_{d\in D}\{P_d g^*\} && (2.3)\\ g^* + v &= \max_{d\in E}\{r_d + P_d v\} && (2.4) \end{aligned}$$

Here, $E \triangleq \{d \in D \mid P_d g^* = g^*\}$ is the set of decision rules which attain the maximum in (2.3). The vector v belongs to a closed, unbounded set of solutions denoted by V. For a given $v \in V$, we denote $E(v) \triangleq \{d \in E \mid g^* + v = L_d v\}$. When the entries of g^* are state-independent, (2.3) and (2.4) reduce to

$$g^* + v = Lv$$

and $E(v)$ is the set of v-improving decision rules.

Our analysis in this paper will refer to combinations of the following conditions on Ψ,

(C1) The optimal gain g^* is state-independent.

(C2) V is one dimensional, i.e. for some v_0, $V = \{v \in \mathbb{R}^{|S|} \mid v = v_0 + c\mathbf{1}, c \in \mathbb{R}\}$

(C3) The sequence $L^k x - kg^*$ converges, as k tends to infinity, for all vectors x.

Perhaps the most general, commonly considered class of models in which (C1) holds is the class of weakly communicating models (see [8, pages 348-352]). The MDP Ψ is weakly communicating if any state is either (i) reachable from all states for some P_d, or (ii) transient for all P_d.

Condition (C2) is also sufficient for (C1). An equivalent condition for (C2), in terms of the chain structure of Ψ, was identified in [9]. Namely, (C2) holds if and only if a randomized optimal decision rule exists with a single recurrent class and this class contains the largest number of recurrent states possible for gain optimal decision rules. A typical case where (C2) holds is when Ψ is a unichain model, meaning that P_d has a single recurrent class for all $d \in D$.

Necessary and sufficient conditions for (C3) were established in [10]. The equivalent condition is that a randomized, aperiodic optimal decision rule exists with the maximum number of recurrent states and the minimum number of recurrent subchains possible for gain optimal decision rules. Typically, (C3) will hold if P_d is aperiodic for all $d \in D$.

If (C3) holds, the operator

$$\hat{L}^{\infty} x \stackrel{\triangle}{=} \lim_{k} (L^k x - k g^*)$$

is defined on all $\mathbb{R}^{|S|}$. It is known that $\hat{L}^{\infty} x \in V$ for all vectors x for which $\hat{L}^{\infty}$ is defined (see Lemma 2.2(g) in [11]). Furthermore $\hat{L}^{\infty}$ has the same properties (in particular, properties (i) and (ii) in Section 2.2) as L (see Lemma 2.1 in [11]).

When (C3) holds in a DSM, we define y_{∞} as the unique vector satisfying the fixed point equations

$$y_{\infty}(s) = \max_{a^{\sigma} \in A^{\sigma}} \{\hat{L}^{\infty} L^{a^{\sigma}} y_{\infty}(s)\}, \quad s \in S \tag{2.5}$$

The uniqueness of y_{∞} follows by noting that the operator applied to y_{∞} on the right hand side is a contraction operator with respect to $||\cdot||$ with contraction factor λ.

In addition, we define

$$M_{y_{\infty}}(\epsilon) \stackrel{\triangle}{=} \{\min k \mid ||L^k L^{a^{\sigma}} y_{\infty} - k g^* - \hat{L}^{\infty} L^{a^{\sigma}} y_{\infty}|| \leq \epsilon \text{ for all } a^{\sigma} \in A^{\sigma}\}.$$

By Lemma 2.2(e) in [11], if (C1) holds, the sequence $||L^k x - k g^* - \hat{L}^{\infty} x||$ decreases monotonically in k. Hence, $M_{y_{\infty}}(\epsilon)$ is the smallest k for which all sequences $L^k L^{a^{\sigma}} y_{\infty}$, $a^{\sigma} \in A^{\sigma}$ have converged within ϵ. Note that $M_{y_{\infty}}(\epsilon)$ is always finite since A^{σ} is finite.

3 Optimality Criteria

In this Section, we describe the discounted optimality criteria which we will work with.

Due to the analogy between our optimality problem and a standard discounted MDP, a theory describing the properties and computation of its solution may be readily derived using standard embedding and reformulation techniques. The details are tangential to our analysis and therefore we merely state the results. Interested readers are referred to the Appendix for relevant proofs and an extended discussion.

3.1 The Discounted PSMDP

Consider a PSMDP and let $0 \leq \lambda < 1$. For a policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}_N^{HR}$, define the following value function

$$v_N^{\boldsymbol{\pi}}(s) \stackrel{\triangle}{=} E_s^{\boldsymbol{\pi}} \textstyle\sum_{t=0}^{\infty} \lambda^{\lfloor t/(N+1) \rfloor} \boldsymbol{r_t}, \quad s \in S \ . \tag{3.1}$$

where $\boldsymbol{r_t}$ is the stochastic process of rewards induced by $\boldsymbol{\pi}$.

We address the optimization problem of finding, for all $s \in S$, the optimal value

$$v_N^*(s) \stackrel{\triangle}{=} \sup_{\boldsymbol{\pi} \in \boldsymbol{\Pi}_N^{HR}} v_N^{\boldsymbol{\pi}}(s) \tag{3.2}$$

as well as a policy, $\boldsymbol{\pi}_{\boldsymbol{\epsilon}}^* \in \boldsymbol{\Pi}_N^{HR}$, satisfying

$$v_N^*(s) - v_{\boldsymbol{\pi}_{\boldsymbol{\epsilon}}^*}(s) \leq \epsilon \tag{3.3}$$

for a given $\epsilon \geq 0$. Policies $\boldsymbol{\pi}_{\boldsymbol{\epsilon}}^*$ satisfying (3.3) are said to be ϵ-optimal. A 0-optimal policy is said to be optimal.

Equations (3.2) and (3.3) define a discounted reward optimality criterion in which the rewards devalue by λ at each renewal epoch. As an example, one may consider a financial operation in which epochs last for minutes or hours but in which renewal cycles last for months or years. In such cases, the present value of the rewards will not change greatly after a single epoch, but significant devaluation may occur on the time scale of renewal cycles.

In the analysis of this paper, we shall treat this optimality problem in detail for DSMs only. The dynamics of an MSM are largely the same as a DSM in which A^σ is a singleton, making extensions of our analysis to MSMs readily obtainable. The extensions are summarized in Section 7.

The following Theorems characterize solutions of the optimality problem. Theorem 3.1 establishes the existence of optimal cyclo-stationary policies and Theorem 3.2 establishes an optimality equation. For proofs, see the Appendix.

Theorem 3.1 (Existence of Cyclo-stationary policies) *An optimal cyclo-stationary policy exists.*

Theorem 3.2 (DSM optimality equation)

(a) The operator $\mathcal{L}_N$ is a contraction operator, with respect to $||\cdot||$, with contraction factor λ and fixed point v_N^*. Hence,

$$v_N^*(s) = \max_{a^\sigma \in A^\sigma} \{L^N L^{a^\sigma} v_N^*(s)\}, \quad s \in S \ . \tag{3.4}$$

(b) A cyclo-stationary policy is optimal if and only if, for each $s \in S$ a sequence $\{a^\sigma, \pi, d\}$ is prescribed satisfying

$$v_N^*(s) = L_\pi^N L_d^{a^\sigma} v_N^*(s). \tag{3.5}$$

3.2 Uniform Optimality

For cyclo-stationary policies in DSMs, we also define the stronger notion of *uniform optimality*. (For MSMs, an analogous definition may be made.) An ϵ-optimal cyclo-stationary policy $\boldsymbol{\pi}_\epsilon^* \in \mathbf{\Pi}_N^{\text{cyc}}$ shall be called *uniform* ϵ-optimal if it is ϵ-optimal starting from *any* time epoch (and not just the initial one). More formally, if a sequence $\{a^\sigma, \pi, d\}$ is prescribed by $\boldsymbol{\pi}_\epsilon^*$ in some initial state $s \in S$,

$$||L^k L^{a^\sigma} v_N^* - L_\pi^k L_d^{a^\sigma} v_N^{\boldsymbol{\pi}_\epsilon^*}|| \leq \epsilon \tag{3.6}$$

for all $0 \leq k < N$. Note that $L_\pi^k L_d^{a^\sigma} v_N^{\boldsymbol{\pi}_\epsilon^*}(s)$ gives the value of $\boldsymbol{\pi}_*^\epsilon$ starting at epoch $N-k$ in state s.

The following obvious result establishes that uniform optimal cyclo-stationary policies exist and relates them to v_N^*.

Theorem 3.3 *A cyclo-stationary policy is uniform optimal if and only if for each $s \in S$ a sequence $\{a^\sigma, \pi, d\}$, is prescribed satisfying,*

$$v_N^*(s) = L_\pi^N L_d^{a^\sigma} v_N^*(s) \tag{3.7}$$

and

$$L^k L^{a^\sigma} v_N^* = L_\pi^k L_d^{a^\sigma} v_N^*, \quad 0 \leq k < N \ . \tag{3.8}$$

The following Lemma provide sufficient conditions, in terms of v_N^*, for cyclo-stationary policies to be ϵ-optimal and uniform ϵ-optimal.

Lemma 3.4 (Criteria for ϵ-optimality)

(a) Suppose that for each fixed $s \in S$, the sequence $\{a^\sigma, \pi, d\}$ prescribed by a cyclo-stationary policy when starting in state s, satisfies

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

Then the cyclo-stationary policy is ϵ-optimal.

(b) Suppose, in addition, that the sequence $\{a^\sigma, \pi, d\}$ satisfies

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

for all $0 \leq k < N$. Then the cyclo-stationary policy is uniform ϵ-optimal.

Uniform (ϵ-)optimal policies are those obtained by using dynamic programming to compute the right hand side of (3.4). In some cases, ϵ-optimal policies of a particularly simple form exist, but which are not uniform optimal. The practical disadvantage of such policies is that they cannot be identified by dynamic programming alone (see [6], Example 8.18 or [7], Example 7.1). Moreover, under (non-uniform) ϵ-optimal policies, with $\epsilon > 0$, a decision maker might reach a state where the sub-optimality (of the reward-to-go) is larger than ϵ (see [6], Example 6.1). In such scenarios, decision makers might wish to deviate from the policy.

4 The Asymptotic Behavior of v_N^*

In this section, we describe the asymptotic behavior of v_N^* as N becomes large. The main result is the following theorem, which establishes that $v_N^* - \frac{Ng^*}{1-\lambda}$ converges to y_∞ geometrically.

Theorem 4.1 (Asymptotic behavior of v_N^*) *Assume (C1) and (C3) hold and let y_∞ be defined as in (2.5). Fix $\epsilon > 0$ and suppose that $N \geq M_{y_\infty}(\epsilon)$. Then*

$$v_N^* = \frac{Ng^*}{1-\lambda} + y_\infty + \bar{o}\left(\frac{\epsilon}{1-\lambda}\right). \tag{4.1}$$

Proof.

Let $z_0 = \frac{Ng^*}{1-\lambda} + y_\infty$. Hence, for arbitrary $a^\sigma \in A^\sigma$,

$$\begin{aligned} L^N L^{a^\sigma} z_0 &= L^N L^{a^\sigma} y_\infty + \frac{\lambda N g^*}{1-\lambda} \\ &= \hat{L}^\infty L^{a^\sigma} y_\infty + Ng^* + \bar{o}(\epsilon) + \frac{\lambda N g^*}{1-\lambda} \\ &= \frac{Ng^*}{1-\lambda} + \hat{L}^\infty L^{a^\sigma} y_\infty + \bar{o}(\epsilon). \end{aligned}$$

The first equality above followed from (C1). The second equality followed from our assumption that $N \geq M_{y_\infty}(\epsilon)$.

Taking maximums over A^σ and noting (2.5), we have,

$$\mathcal{L}_N z_0 = \frac{Ng^*}{1-\lambda} + y_\infty + \bar{o}(\epsilon) = z_0 + \bar{o}(\epsilon).$$

By Theorem 3.2, the Banach theorem then implies,

$$||v_N^* - z_0|| \leq \frac{||\mathcal{L}_N z_0 - z_0||}{1-\lambda} \leq \frac{\epsilon}{1-\lambda}$$

from which (4.1) follows. $\square$

Remark 4.2 The Theorem implies that for any $\epsilon > 0$

$$||v_N^* - \frac{Ng^*}{1-\lambda} - y_\infty|| \leq \epsilon$$

for $N \geq M_{y_\infty}((1-\lambda)\epsilon)$. This means that the convergence of this sequence is as fast as some non-discounted value iteration sequence, $L^k x - kg^*$ i.e. it is geometric (See [11]).

When (C2) holds, we obtain additional information about the vector y_∞, namely that $y_\infty \in V$. We show this formally in the following Theorem.

Theorem 4.3 (Properties of y_∞) *Assume (C2) and (C3) hold. Then in the fixed point equation*

$$y_\infty(s) = \max_{a^\sigma \in A^\sigma} \{\hat{L}^\infty L^{a^\sigma} y_\infty(s)\}, \quad s \in S$$

a single slowscale action a_*^σ attains the maximum for all $s \in S$. Hence,

$$y_\infty = \hat{L}^\infty L^{a_*^\sigma} y_\infty \tag{4.2}$$

and $y_\infty \in V$.

Proof.

Since $\hat{L}^\infty$ maps into V, (C2) implies the representation

$$\hat{L}^\infty L^{a^\sigma} y_\infty = v_0 + c^{a^\sigma} \mathbf{1} \tag{4.3}$$

for all $a^\sigma \in A^\sigma$. Choose a_*^σ so that $\max_{a^\sigma \in A^\sigma} \{c^{a^\sigma}\} = c^{a_*^\sigma}$ and fix an arbitrary $s \in S$. Then

$$\begin{aligned} y_\infty(s) &= \max_{a^\sigma \in A^\sigma} \{\hat{L}^\infty L^{a^\sigma} y_\infty(s)\} \\ &= \max_{a^\sigma \in A^\sigma} \{v_0(s) + c^{a^\sigma}\} \\ &= v_0(s) + c^{a_*^\sigma} \end{aligned} \tag{4.4}$$

Since s was arbitrary, this implies that $y_\infty = v_0 + c^{a_*^\sigma} \mathbf{1}$. Equation (4.2) follows from this result and (4.3) with $a^\sigma = a_*^\sigma$. In addition, (4.2) implies $y_\infty \in V$ since $\hat{L}^\infty$ maps into V. □

5 Approximate Optimality of Initially Stationary Policies

In this section, we establish the existence of ϵ-optimal i.s.p.'s, for N sufficiently large, when the underlying MDP has state-independent gain.

Our basic result, Theorem 5.1, indicates that an ϵ-optimal i.s.p. exists whose planning horizon η_ϵ is bounded independently of N. In addition, the initial decision rules are gain optimal in Ψ.

When (C3) holds, the right hand side (3.4) gives us some reason to anticipate such a result mathematically. If the sequences $\{L^k L^{a^\sigma} v_N^* - kg^*,\ a^\sigma \in A^\sigma\}$ have nearly converged after $\eta_\epsilon < N$ steps, then $\{L^\eta L^{a^\sigma} v_N^*,\ a^\sigma \in A^\sigma\}$ approximate elements of V and, for each $a^\sigma \in A^\sigma$, a $\delta \in D^*$ exists satisfying

$$L_\delta^{N-\eta_\epsilon} L^{\eta_\epsilon} L^{a^\sigma} v_N^* \approx L^N L^{a^\sigma} v_N^*. \tag{5.1}$$

Hence, decision sequences corresponding to an i.s.p. with planning horizon η_ϵ are approximately maximizing in the PSMDP optimality equation.

For (5.1) to be satisfied, we might initially expect to require some condition which bounds the rate of convergence of the sequences $\{L^k L^{a^\sigma} v_N^* - kg^*,\ a^\sigma \in A^\sigma\}$. Otherwise, since v_N^* is a priori unknown, we cannot be sure that convergence will take place in $\eta_\epsilon < N$ steps. A virtue of our subsequent analysis is that it does not require any such conditions. Apart from the state-independence of g^*, Theorem 5.1 requires only condition (C3), the existence of the limit. For Theorem 5.2, we add condition (C2) and prove the existence of ϵ-optimal i.s.p.'s of a simplified structure. Here again, however, no information about the rate of convergence of value iteration is used.

Condition (C3) is a technical assumption which excludes periodic phenomenon, thereby simplifying our analysis. When (C3) does not hold, it is known (see [10]) that the sequences $L^k x - kg^*$ are asymptotically periodic. The above arguments then generalize, establishing that ϵ-optimal cyclo-stationary policies exist which use a periodically varying sequence of decision rules at the beginning of each renewal cycle.

Theorem 5.1 (Existence of ϵ-optimal i.s.p.'s) *Assume (C1) and (C3) hold. Fix $\epsilon > 0$ and let $\eta_\epsilon = M_{y_\infty}((1-\lambda)^2\epsilon/2)$.*

Then if $N > \eta_\epsilon$, a uniform ϵ-optimal initially stationary policy with planning horizon η_ϵ exists. The i.s.p. may be chosen so that for each initial state $s \in S$,

(i) A slowscale action $a^\sigma \in A^\sigma$, which is optimal in s for some optimal cyclo-stationary policy, is selected.

(ii) The initial decision rule for s is $\hat{L}^\infty L^{a^\sigma} y_\infty$-improving where a^σ is as in (i).

Proof.

Consider an initial state $s \in S$. Fix a slowscale action $a^\sigma \in A^\sigma$ which is chosen in s by some optimal cyclo-stationary policy. Let δ be any $\hat{L}^\infty L^{a^\sigma} y_\infty$-improving decision rule, let $d \in D$ satisfy $L^{a^\sigma} y_\infty = L_d^{a^\sigma} y_\infty$, and let π be a sequence of N decision rules whose first $N - \eta_\epsilon$ terms are δ, and which satisfies $L^k L^{a^\sigma} y_\infty = L_\pi^k L_d^{a^\sigma} y_\infty$ for $0 \le k \le \eta_\epsilon$.

Since $N > \eta_\epsilon = M_{y_\infty}((1-\lambda)^2\epsilon/2)$, we have from Theorem 4.1 that

$$v_N^* = \frac{Ng^*}{1-\lambda} + y_\infty + \bar{o}((1-\lambda)\epsilon/2).$$

In light of (C1), the last equation implies,

$$||L^k L^{a^\sigma} v_N^* - L_\pi^k L_\pi^{a^\sigma} v_N^*|| \leq ||L^k L^{a^\sigma} y_\infty - L_\pi^k L_d^{a^\sigma} y_\infty|| + \lambda(1-\lambda)\epsilon. \tag{5.2}$$

for all $0 \leq k \leq N$.

For $0 \leq k \leq \eta_\epsilon$, the definition of π implies that the first term on the right hand side of (5.2) is zero. Therefore, we deduce the bound

$$||L^k L^{a^\sigma} v_N^* - L_\pi^k L_d^{a^\sigma} v_N^*|| \quad \leq \quad (1-\lambda)\epsilon. \tag{5.3}$$

Furthermore, by our definition of η_ϵ, we have that, for $k \geq \eta_\epsilon$,

$$L^k L^{a^\sigma} y_\infty = \hat{L}^\infty L^{a^\sigma} y_\infty + kg^* + \bar{o}((1-\lambda)^2\epsilon/2). \tag{5.4}$$

In the particular case where $k = \eta_\epsilon$,

$$L^{\eta_\epsilon} L^{a^\sigma} y_\infty = \hat{L}^\infty L^{a^\sigma} y_\infty + \eta_\epsilon g^* + \bar{o}((1-\lambda)^2\epsilon/2). \tag{5.5}$$

Hence, for $k > \eta_\epsilon$,

$$\begin{aligned} L_\pi^k L_d^{a^\sigma} y_\infty &= L_\delta^{k-\eta_\epsilon}(L^{\eta_\epsilon} L^{a^\sigma} y_\infty) \\ &= \hat{L}^\infty L^{a^\sigma} y_\infty + kg^* + \bar{o}((1-\lambda)^2\epsilon/2). \end{aligned} \tag{5.6}$$

where the last inequality combines (5.5) with the fact that δ is $\hat{L}^\infty L^{a^\sigma} y_\infty$-improving and $\hat{L}^\infty L^{a^\sigma} y_\infty \in V$. Combining (5.4) and (5.6), we deduce that

$$||L^k L^{a^\sigma} y_\infty - L_\pi^k L_d^{a^\sigma} y_\infty|| \leq (1-\lambda)^2\epsilon$$

for $\eta_\epsilon < k \leq N$ and so (5.2) becomes

$$\begin{aligned} ||L^k L^{a^\sigma} v_N^* - L_\pi^k L_d^{a^\sigma} v_N^*|| &\leq (1-\lambda)^2\epsilon + \lambda(1-\lambda)\epsilon \\ &= (1-\lambda)\epsilon \end{aligned} \tag{5.7}$$

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

Finally, (5.3) and (5.7) together imply that

$$||L^k L^{a^\sigma} v_N^* - L_\pi^k L_d^{a^\sigma} v_N^*|| \leq (1-\lambda)\epsilon \tag{5.8}$$

for $0 \leq k \leq N$. For the particular case $k = N$, (5.8) becomes

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

by our choice of a^σ and Theorem 3.2.

By Lemma 3.4, a sequence $\{a^\sigma, \pi, d\}$ constructed for all initial states in this manner corresponds to a uniform ϵ-optimal i.s.p as described in the statement of the Theorem. □

When (C2) holds, we obtain a simplification of Theorem 5.1. It indicates the existence of an ϵ-optimal i.s.p. which prescribes actions based on at most the current state and time (i.e. it is Markovian deterministic). Since DSMs have history dependent transition probability and reward functions, this result is not self-evident.

Theorem 5.2 (Existence of Markovian i.s.p.'s) *Assume (C2) and (C3) hold. Fix $\epsilon > 0$ and let $\eta_\epsilon = M_{y_\infty}((1-\lambda)^2\epsilon/6)$.*

Then if $N > \eta_\epsilon$, a uniform ϵ-optimal initially stationary policy with planning horizon η_ϵ exists. The i.s.p. can be constructed so that

(i) It prescribes the same sequence $\{a_*^\sigma, \pi, d\}$ in each initial state where a_*^σ is as in Theorem 4.3 (Consequently it is Markovian deterministic).

(ii) The initial decision rule is v_0-improving.

Proof.

Let δ be a v_0-improving decision rule and let $a_*^\sigma \in A^\sigma$ be defined as in Theorem 4.3. Note that, since $\hat{L}^\infty$ maps into V, (C2) implies that δ is also $\hat{L}^\infty L^{a^\sigma} y_\infty$-improving for any $a^\sigma \in A^\sigma$.

In addition, let $d \in D$ be a decision rule satisfying $L^{a_*^\sigma} y_\infty = L_d^{a_*^\sigma} y_\infty$ and let π be a sequence of N decision rules π whose first $N - \eta_\epsilon$ terms are δ and which satisfies $L^k L^{a_*^\sigma} y_\infty = L_\pi^k L_d^{a_*^\sigma} y_\infty$ for $0 \le k \le \eta_\epsilon$.

Using the same manipulations as in the proof of Theorem 5.1, with ϵ replaced by $\epsilon/3$, we find, as in (5.4), that

$$L^k L^{a_*^\sigma} y_\infty = \hat{L}^\infty L^{a_*^\sigma} y_\infty + kg^* + \bar{o}((1-\lambda)^2\epsilon/6) \tag{5.9}$$

for all $k \ge \eta_\epsilon$. Also, as in (5.8), we find that

$$||L^k L^{a_*^\sigma} v_N^* - L_\pi^k L_d^{a_*^\sigma} v_N^*|| \le (1-\lambda)\epsilon/3 \tag{5.10}$$

for all $0 \le k \le N$. Noting Theorem 4.3 and letting $k = N$, (5.9) becomes

$$L^N L^{a_*^\sigma} y_\infty = y_\infty + Ng^* + \bar{o}((1-\lambda)^2\epsilon/6).$$

Also, by Theorem 4.1,

$$v_N^* = \frac{Ng^*}{1-\lambda} + y_\infty + \bar{o}((1-\lambda)\epsilon/6).$$

The last two equations can be used to compute $v_N^* - L^N L^{a_*^\sigma} v_N^*$ by direct substitution. Noting that g^* is state-independent, this leads to

$$||v_N^* - L^N L^{a_*^\sigma} v_N^*|| \le (1-\lambda)2\epsilon/3. \tag{5.11}$$

Therefore, using the triangle inequality,

$$\begin{aligned}
&||v_N^* - L_\pi^N L_d^{a_*^\sigma} v_N^*|| \\
&\leq \; ||v_N^* - L^N L^{a_*^\sigma} v_N^*|| + ||L^N L^{a_*^\sigma} v_N^* - L_\pi^N L_d^{a_*^\sigma} v_N^*|| \\
&\leq \; (1-\lambda)\epsilon
\end{aligned}$$

where the second inequality incorporated (5.10) with $k = N$ and (5.11). Hence,

$$v_N^*(s) - L_\pi^N L^{a_*^\sigma} v_N^*(s) \leq (1-\lambda)\epsilon$$

for all $s \in S$.

This, together with (5.10) and Lemma 3.4 implies the result. □

The Theorem entails a significant simplification. Since the ϵ-optimal i.s.p. which it describes uses the same sequence in every state, only $O(\eta_\epsilon |S|)$ data are needed to represent it, as opposed to the $O(N|S||A^\sigma|)$ data typically required by cyclo-stationary policies (see Remark A.2, in the Appendix) or the $O(\eta|S||A^\sigma|)$ data typically required by i.s.p.'s.

Procedures for constructing ϵ-optimal i.s.p.'s as described in Theorems 5.1 and 5.2 are outlined in [6], Section 7.6 and 7.7. These Procedures follow from the constructive nature of the proofs of these Theorems. Mainly, what is involved is finding y_∞ and evaluating the dynamic programming sequences $\{L^k L^{a^\sigma} y_\infty, \, a^\sigma \in A^\sigma\}$. Notably, when (C2) holds, finding y_∞ is effectively the same as solving the underlying MDP, due to Theorem 4.3.

6 Optimal Policies

In this Section, we assume (C2) and (C3) and examine the structure of *precisely* optimal policies for large N.

With v_0 as in (C2), let

$$\rho \overset{\triangle}{=} \frac{1}{2} \min_{d \in D \setminus E(v_0)} ||Lv_0 - L_d v_0|| \tag{6.1}$$

with $\rho = \infty$ if $E(v_0) \neq D$. The parameter ρ satisfies

$$\operatorname*{argmax}_{d \in D} \{r_d + P_d v_0 + \bar{o}(\rho)\} \subset E(v_0). \tag{6.2}$$

The following Theorem establishes that uniform optimal policies use v_0-improving decision rules at all epochs but some final planning horizon in each renewal cycle.

Theorem 6.1 (Structure of optimal policies) *Assume (C2) and (C3) hold and let* $\eta_0 = M_{y_\infty}((1-\lambda)\rho)$. *Then if* $N > \eta_0$*, all uniform optimal cyclo-stationary policies prescribe decision rules in* $E(v_0)$ *on the first* $N - \eta_0$ *fastscale epochs of each renewal cycle.*

Proof.

Consider an arbitrary uniform optimal cyclo-stationary policy and initial state $s \in S$. Denote the sequence prescribed by π in s as $\{a^\sigma, \pi, d\}$.

Since $N > \eta_0$,

$$v_N^* = \frac{Ng^*}{1-\lambda} + y_\infty + \bar{o}(\rho)$$

by Theorem 4.1.

Also, by our definition of η_0 we have,

$$L^k L^{a^\sigma} y_\infty = \hat{L}^\infty L^{a^\sigma} y_\infty + kg^* + \bar{o}((1-\lambda)\rho)$$

for all $k \geq \eta_0$.

Noting that g^* is state-independent, the last two equations may be combined to obtain

$$\begin{aligned} L^k L^{a^\sigma} v_N^* &= L^k L^{a^\sigma} y_\infty + \frac{\lambda N g^*}{1-\lambda} + \lambda \bar{o}(\rho) \\ &= \hat{L}^\infty L^{a^\sigma} y_\infty + \left(k + \frac{\lambda N}{1-\lambda}\right) g^* + \bar{o}(\rho). \end{aligned}$$

for all $k \geq \eta_0$. However $\hat{L}^\infty L^{a^\sigma} y_\infty \in V$ and so, by (C1) and (C2), the expression $\hat{L}^\infty L^{a^\sigma} y_\infty + \left(k + \frac{\lambda N}{1-\lambda}\right) g^*$ differs from v_0 by a state-independent vector. Incorporating this into the preceding equality and recalling (6.2), we have

$$\operatorname*{argmax}_{d \in D}\{r_d + P_d(L^k L^{a^\sigma} v_N^*)\} = \operatorname*{argmax}_{d \in D}\{r_d + P_d v_0 + \bar{o}(\rho)\} \subset E(v_0)$$

for all $k \geq \eta_0$.

Since the cyclo-stationary policy is uniform optimal, this last result implies, by Theorem 3.3, that all but the final η_0 terms of π are v_0-improving. Since the cyclo-stationary policy and initial state considered were arbitrary, the assertions of the Theorem are proven. □

Theorem 6.1 has the following Corollary:

Corollary 6.2 (Existence of optimal i.s.p.'s) *Assume (C2) and (C3) hold and let* $\eta_0 = M_{y_\infty}((1-\lambda)\rho)$. *Then if* $E(v_0)$ *is a singleton,* δ*, and* $N > \eta_0$*, a uniform optimal simple i.s.p. exists with planning horizon* η_0 *and initial decision rule rule* δ*.*

Proof.

Immediate from the Theorem. □

Example 7.4 in [6] provides a counter-example for the Corollary. The requirement of the Corollary that there exists a unique v_0-improving decision rule is not a condition which has been widely analyzed in the literature, to the best of our knowledge. Remark 3 of [11] refers to the case where D^* is a singleton, which in turn implies that $E(v_0)$ is a singleton since $E(v_0) \subset D^*$. However, no assessment of the likelihood of this condition is provided.

In [6], a certain measure of the likelihood of uniqueness was made in a simplified setting. Data for 100,000 2-state models were randomly generated according to a uniform distribution and the instances of non-uniqueness were counted. A 0% incidence of non-uniqueness was found giving at least some indication that uniqueness might be the the generic case.

Procedure 7.16 in [6] is a policy iteration algorithm for finding the optimal i.s.p. described by Corollary 6.2.

7 Extension to Markovian Slowscale Models (MSMs)

All of our analysis for DSMs in the preceding sections followed from the DSM optimality equation (3.4) together with the related dynamic programming sequences $\{L^k L^{a^\sigma} v_N^*,\ a^\sigma \in A^\sigma\}$.

By applying reasoning analogous to that in the Appendix to an MSM, one obtains the optimality equation

$$v_N^* = L^N L^\sigma v_N^*. \tag{7.1}$$

Furthermore, we find that optimality is obtained on a class of Markovian deterministic policies, consisting of periodically repeating decision rule sequences of the form

$$\{d_0, d_1, \ldots, d_{N-1}, d^\sigma\} \tag{7.2}$$

where $d_i \in D$, $i = 0, 1, \ldots N-1$ and $d^\sigma \in D^\sigma$.

An analysis analogous to that in the preceding sections can therefore be carried out by considering (7.1) and the dynamic programming sequences $L^k L^\sigma v_N^*$. This analysis is essentially the same as that of a DSM whose slowscale action space consists of a single action, except that the role of the single L^{a^σ} operator is assumed by L^σ.

Consequently, conclusions about the behavior of v_N^* as a function of N and the existence of i.s.p.'s can be drawn by applying results already derived to this special case. So, for example, we deduce that there exist policies of the form

$$\{\delta, \delta, \delta, \ldots, \delta, d_{N-\eta}, d_{N-\eta+1}, \ldots, d_{N-1}, d^\sigma\} \tag{7.3}$$

which are ϵ-optimal under the conditions of Theorem 5.1 and precisely optimal under the conditions of Corollary 6.2. Likewise, the relationship of the initial decision rules δ and planning horizon η to Ψ, are as described in those Theorems.

The conclusions of Theorem 5.2 imply no special simplification in this context because the existence of optimal Markovian policies is trivial for MSMs.

8 On the Extension to the State-Dependent Gain Case

The previous sections complete the analysis of DSMs where the underlying MDP has state-independent optimal gain. In this Section, we make a preliminary examination of the state-dependent gain case to illustrate its distinctiveness from state-independent gain models. This is in preparation for a sequel paper [7] where a fuller analysis of the state-dependent gain case will be given.

In the state-independent gain case, we found that the structure of ϵ-optimal policies were influenced by the limiting behavior of the sequences $\{L^k L^{a^\sigma} v_N^* - kg^*,\, a^\sigma \in A^\sigma\}$. These sequences nearly converge in a bounded number of steps. Recalling the discussion in Section 5, ϵ-optimal i.s.p.'s therefore exist with gain optimal initial decision rules.

Theorem 8.1 provides some alternative insight as to why this occurs in the state-independent gain case, and also as to why more complicated behavior is apt to be encountered when g^* is state-dependent. The Theorem establishes that the optimal value v_N^* is bounded as a function of N (in span norm) if and only if g^* is state-independent. Otherwise, v_N^* diverges linearly in N. Note that this is the fastest rate at which it could diverge.

Since $||v_N^*||_{\text{sp}}$ is bounded when $g^*(\text{s})$ is state-independent, the sequences $\{L^k L^{a^\sigma} v_N^* - kg^*,\, a^\sigma \in A^\sigma\}$ are equivalent to value iteration sequences on a bounded set of terminal rewards. Conditions on state-independent average reward models are known which bound the rate of convergence of such sequences (e.g. Theorem 2(d) in [12]). Hence, the number of steps for approximate convergence is likewise bounded in N. The Theorem therefore allows one to anticipate the existence of ϵ-optimal i.s.p.'s in the state-independent gain case under established conditions on the rate of convergence of value iteration. Once again, however, our analysis did not rely on such conditions.

Conversely, when g^* is state-dependent, $||v_N^*||_{\text{sp}}$ is unbounded in N. Hence, the number of steps for approximate convergence will, in general, be unbounded as well, possibly even greater than N. In the former case, ϵ-optimal i.s.p.'s might require planning horizons which grow as an unbounded function of N. This is illustrated in Example 8.3. In the latter case, it means that the structure of ϵ-optimal policies is not determined by the limiting behavior of value iteration at all. Example 8.4 shows how, in the absence of appropriate conditions on the stationary data, the non-stationary data can affect whether the sequences $\{L^k L^{a^\sigma} v_N^* - kg^*,\, a^\sigma \in A^\sigma\}$ approximately converge in $\eta < N$ steps. When this convergence does not take place, optimal i.s.p.'s with non-gain optimal initial decision rules may be observed, in contrast to our previous results.

Theorem 8.1 (Boundedness of v_N^* in N)

(a) $||g^*||_{\mathrm{sp}} = 0 \Rightarrow \limsup_{N\to\infty} ||v_N^*||_{\mathrm{sp}} < \infty$.

(b) $||g^*||_{\mathrm{sp}} > 0 \Rightarrow \liminf_{N\to\infty} \frac{||v_N^*||_{\mathrm{sp}}}{N} > 0$.

Proof.

(a) For any vector $x \in \mathbb{R}^{|S|}$,

$$\mathcal{L}_N x(s) = \max_{a^\sigma \in A^\sigma} \{L^N L^{a^\sigma} x(s) - Ng^*(s)\} + Ng^*(s) \tag{8.1}$$

for all $s \in S$.

Let $x \in \mathbb{R}^{|S|}$ be a fixed vector of span 0. By Theorem 9.4.1(a) in [8], with $v_0 = L^{a^\sigma} x$, the first term on the right hand side of (8.1) is bounded with respect to N. Since $||g^*||_{\mathrm{sp}} = 0$, it follows that $||\mathcal{L}_N x||_{\mathrm{sp}}$ is bounded in N.

Now, from the triangle inequality, Theorem 3.4(a), and the Banach Theorem

$$\begin{aligned} ||v_N^*||_{\mathrm{sp}} &\le ||v_N^* - x||_{\mathrm{sp}} + ||x||_{\mathrm{sp}} \\ &\le \frac{||\mathcal{L}_N x - x||_{\mathrm{sp}}}{1-\lambda} + ||x||_{\mathrm{sp}} \\ &= \frac{||\mathcal{L}_N x||_{\mathrm{sp}}}{1-\lambda}. \end{aligned}$$

Since $||\mathcal{L}_N x||_{\mathrm{sp}}$ is bounded in N, part (a) follows.

(b) Aiming for a contradiction suppose $||g^*||_{\mathrm{sp}} > 0$ but that there exists a subsequence N_k along which

$$\lim_k \frac{||v_{N_k}^*||_{\mathrm{sp}}}{N_k} = 0 \tag{8.2}$$

Now let $x = v_{N_k}^*$ in (8.1) and divide through by N_k. Then noting Theorem 3.2(a), equation (8.1) becomes,

$$v_{N_k}^*(s)/N_k = \max_{a^\sigma \in A^\sigma} \{L^N L^{a^\sigma} v_{N_k}^*(s) - N_k g^*(s)\}/N_k + g^*(s), \quad s \in S\ .$$

The terms $v_{N_k}^*(s)/N_k$ and $\max_{a^\sigma \in A^\sigma} \{L^N L^{a^\sigma} v_{N_k}^*(s) - N_k g^*(s)\}/N_k$ are components of vectors whose spans approach zero as k tends to infinity. This is due to (8.2) and to Theorem 9.4.1(a) in [8] with $v_0 = L^{a^\sigma} v_{N_k}^*$. Therefore, passing to the limit, we deduce from the last equation that $||g^*||_{\mathrm{sp}} = 0$, establishing a contradiction. □

Remark 8.2 The boundedness of v_N^* in N depends only on the stationary data and not on the non-stationary data.

Example 8.3 (Unbounded planning horizons) Consider a DSM with

$$S = \{1, 2\},\ A_1 = \{1, 2\},\ A_2 = \{1\},\ \lambda = .8$$

and stationary data,

$$\begin{array}{ccccc} s & a & r(s,a) & p(1|s,a) & p(2|s,a) \\ \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} & \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & \multicolumn{2}{c}{\begin{bmatrix} 1 & 0 \\ 0.5 & 0.5 \\ 0 & 1 \end{bmatrix}} \end{array}$$

In addition, $A^\sigma = \{a^\sigma\}$ and,

$$r_d^{a^\sigma} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},\quad P_d^{a^\sigma} = \begin{bmatrix} 0 & 1 \\ .75 & .25 \end{bmatrix}$$

for all $d \in D$.

We shall let $\delta \in D$ denote the decision rule which chooses action 1 in state 1 while $\zeta \in D$ will denote the decision rule which chooses action 2. Also, we shall use the notation $\text{seg}_2(d_1, d_2, \eta),\ 0 \le \eta \le N$ to denote the "two-segment" i.s.p. which uses $d_1 \in D$ for the first $N - \eta$ epochs of each renewal cycle and $d_2 \in D$ for the remaining ones. It is immediate to verify that δ is the only decision rule in D^* (cf. Section 2.4) and that $g^* = [\ 1 \quad 0\ ]^T$. Hence (C1) does not hold. Since all decision rules are aperiodic (C3) does hold.

The optimality equation is $v_N^* = L^N L^{a^\sigma} v_N^*$. Since state 2 is absorbing with reward zero on the stationary epochs, the optimality equation for state 2 is

$$v_N^*(2) = 0.6 v_N^*(1) + 0.2 v_N^*(2).$$

Hence v_N^* has the form $[\ c_N^* \quad 0.75 c_N^*\ ]^T$. Since positive rewards are always obtainable in state 1, $c_N^* > 0$. Optimal policies are obtained via the backward induction sequence,

$$\begin{bmatrix} c_N^* \\ 0.75 c_N^* \end{bmatrix} = L^N \begin{bmatrix} 0.6 c_N^* \\ 0.75 c_N^* \end{bmatrix}$$

Initially, ζ is a maximizing decision rule in this sequence. If at some stage δ is maximizing, it remains maximizing. Therefore, a policy of the form $\text{seg}_2(\delta, \zeta, \eta)$ is optimal.

Imitating the above, the value of $\text{seg}_2(\delta, \zeta, \eta)$ can be expressed as $[\ c_N^\eta \quad 0.75 c_N^\eta\ ]^T$ and satisfies the fixed point equation,

$$\begin{bmatrix} c_N^\eta \\ 0.75 c_N^\eta \end{bmatrix} = L_\delta^{N-\eta} L_\zeta^\eta L^{a^\sigma} \begin{bmatrix} c_N^\eta \\ 0.75 c_N^\eta \end{bmatrix}.$$

Solving the component equation for state 1, we obtain

$$c_N^\eta = \frac{N-\eta}{0.25+0.15(2^{-\eta})}. \tag{8.3}$$

The optimality problem therefore reduces to finding $\eta^*(N)$ which maximizes c_N^η. Optimizing the right hand side of (8.3) over η, we obtain that $\eta^*(N)$ is of order $\log_2 N$ and $\lim_{N\to\infty} \frac{c_N^*}{N} = 4$.

We now compare the value of the optimal policy $\text{seg}_2(\delta, \zeta, \eta^*(N))$ with that of an initially stationary policy $\text{seg}_2(\delta, \zeta, \eta_0)$ having a fixed planning horizon η_0. It is sufficient to examine $c_N^* - c_N^{\eta_0}$ for which

$$\lim_{N\to\infty} \frac{c_N^* - c_N^{\eta_0}}{N} = 4 - \frac{1}{0.25+0.15(2^{-\eta_0})} > 0.$$

Thus, we see that the discrepancy between the optimal value and the value of the i.s.p. with a bounded planning horizon diverges at a linear rate as N tends to infinity. It is obvious that the same conclusion holds for $\text{seg}_2(\zeta, \delta, \eta_0)$ since such a policy can only accrue non-zero rewards on the planning horizon.

We conclude, therefore, that the size of the planning horizon required in order for an i.s.p. to be ϵ-optimal in this model is an unbounded function of N.

Example 8.4 (Initial decision rules which are not gain optimal) Consider a DSM with

$$S = \{1,2,3\},\ A_1 = \{1,2\},\ A_2 = \{1\},\ A_3 = \{1,2\},\ \lambda = .8$$

and stationary data

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

where the parameter $r > 0$. In addition, $A^\sigma = \{a^\sigma\}$ and

$$r_d^{a^\sigma} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad P_d^{a^\sigma} = \begin{bmatrix} 0 & 1 & 0 \\ .75 & .25 & 0 \\ 1-\mu & 0 & \mu \end{bmatrix}$$

for all $d \in D$. Here $0 < \mu < 1$. We will show that, for N sufficiently large and for certain choices of μ and r, the only (ϵ-)optimal cyclo-stationary policy is an i.s.p. whose initial decision rule is not average optimal. In particular, the initial decision rule chooses action 2 in state 3.

Since state 3 cannot be reached when starting in states 1 or 2, the system reduces to that in Example 8.3 for these starting states. Consequently, $v_N^*(1) = c_N^*, v_N^*(2) = 0.75c_N^*$ and the optimal decisions in states 1 and 2 at all epochs are given by the i.s.p. $\mathrm{seg}_2(\delta, \zeta, \eta^*)$ from Example 8.3 . As for state 3, the optimality equation implies

$$v_N^* = L^N[\ 0.6c_N^* \quad 0.75c_N^* \quad 0.8c_N^*(1-\mu) + 0.8\mu v_N^*(3)\ ]^T.$$

When $N > 1$, it is clear that there are only two sequences of actions which can attain the maxima on the right hand side of the last equation for initial state 3. One possible sequence is to remain in state 3 for all N stationary epochs and thereby obtain terminal reward $0.8c_N^*(1-\mu) + 0.8\mu v_N^*(3)$ at cost rN. The second is to absorb immediately into state 2 and collect a terminal reward of $0.75c_N^*$. The only alternative to these two sequences is to absorb into state 2 after k fastscale epochs where $0 < k < N$. Doing so yields the same terminal reward $0.75c_N^*$ as in the previous case but at cost kr. Hence, it cannot be a maximizing sequence. The optimality equation for state 3 therefore reduces to

$$v_N^*(3) = \max\{0.75c_N^*, 0.8c_N^*(1-\mu) + 0.8\mu v_N^*(3) - rN\}.$$

We are interested in the case where only the second argument on the right hand side attains the maximum. When this is so, staying in state 3 throughout the renewal cycle is the only possibility in an optimal cyclo-stationary policy and

$$v_N^*(3) = \frac{4(1-\mu)c_N^* - 5rN}{5-4\mu}. \tag{8.4}$$

From (8.4), this will be the case when

$$v_N^*(3) - 0.75c_N^* = \frac{(0.25-\mu)c_N^* - 5rN}{5-4\mu} > 0.$$

In our analysis of Example 8.3, we saw that c_N^* is of the order $4N$. Hence, upon normalizing by N and taking N sufficiently large, the last equality becomes, in view of (8.3)

$$\lim_{N\to\infty} \frac{v_N^*(3) - 0.75c_N^*}{N} = \frac{1-4\mu-5r}{5-4\mu} \triangleq \kappa(\mu, r) \tag{8.5}$$

For sufficiently small r and μ, we have $\kappa(\mu, r) > 0$. In this case, action 2 is the only optimal decision in state 3, for N sufficiently large. Also, $N\kappa(\mu, r) > 0$ is approximately the minimum amount by which non-optimal cyclo-stationary policies are suboptimal. Hence ϵ-optimality (or even Nϵ-optimality) is possible for arbitrary $\epsilon \geq 0$ only via an i.s.p. with a non-average optimal initial decision rule. In turn, this indicates that $L^k L^{a^\sigma} v_N^* - kg^*$ does not approximately converge for $k \leq N$.

Conversely, when $\kappa(\mu, r) < 0$, action 1 is always chosen, implying that a uniform optimal i.s.p. exists with a gain optimal initial decision rule. Moreover, $L^k L^{a^\sigma} v_N^* - kg^*$ does approximately converge for $k \leq N$.

9 Conclusions

The results we have presented focus on a discounted optimality problem for PSMDPs whose underlying MDP has state-independent optimal gain. In this case, we arrive at a complete characterization of the optimal value, v_N^*, for large N. In particular, Theorem 4.1 establishes that the sequence $v_n^* - \frac{ng^*}{1-\lambda}$ converges geometrically. In addition, we have found that initially stationary policies possessing favorable properties exist. Given any $\epsilon > 0$, Theorem 5.1 assures that a uniform ϵ-optimal i.s.p. can be found, provided N is sufficiently large. Moreover, the length of the planning horizon has a bound depending only on ϵ and not on N. When N greatly exceeds this bound, the form of the i.s.p. is much simpler than a general cyclo-stationary policy.

These general results rest on fairly weak assumptions. Beyond the assumption of state-independent gain itself, we have applied condition (C3). However, this mainly simplifies the presentation of our analysis. When (C3) does not hold, generalizations of our analysis which account for periodicity effects are possible, although trite.

When, in addition (C2), a condition of common interest in the literature, is applied, a significant simplification is obtained for DSMs. Theorem 5.2 establishes that, for sufficiently large N, a Markovian deterministic uniform ϵ-optimal i.s.p. exists, despite the fact that DSMs are non-Markovian models. Under (C2), we have also established that uniform optimal cyclo-stationary policies use gain optimal decision rules which can be obtained from the underlying MDP's average optimality equations. For sufficiently large N, these decision rules are used on all fastscale epochs apart from some planning horizon which does not depend on N. When only one gain optimal decision rule can be obtained from the average optimality equations, we concluded (see Corollary 6.2) that precisely optimal i.s.p.'s exist.

Of the case where the underlying MDP has state-dependent gain, we have made a preliminary examination. Examples and a general theorem were presented which indicate that the behavior of PSMDPs with state-dependent gain is more complex than those with state-independent gain. A fuller treatment of such models will be given in a sequel paper.

Appendix

This Appendix contains the proofs of the statements in Section 3, and some additional remarks concerning these results.

Proof of Theorem 3.1 (Existence of Cyclo-stationary policies). The discounted DSM can be embedded in a standard, stationary discounted MDP in which rewards are discounted by $\lambda^{1/N}$ at every epoch. To implement this embedding, we consider a state space

$$\hat{S} = S \cup (S \times A^{\sigma} \times \{0, 1, 2, \dots, N\}).$$

where $\times$ denotes the Cartesian product.

Transition probabilities are then chosen such that, with probability one, a state of the form $s \in S$ is occupied at each renewal epoch, while at all other epochs, the state is a triplet of the form

$$(s, \alpha^\sigma, \tau) \in S \times A^\sigma \times \{0, 1, 2, \dots, N\}.$$

In the latter case, s will be the element in S currently observed at t; α^σ will be the particular slowscale action taken at the most recent renewal epoch; and τ will equal $t \bmod N$, the time elapsed since the most recent renewal epoch.

Reward functions are chosen such that

$$\begin{aligned} \hat{r}(s, (a, a^\sigma)) &= r_0(s, a, a^\sigma) \\ \hat{r}((s, a^\sigma, \tau), a) &= \frac{r_\tau(s, a, a^\sigma)}{\lambda^{\tau/(N+1)}} \end{aligned}$$

In this alternative MDP formulation, an optimal stationary deterministic policy exists. It follows by analogy that an optimal $\boldsymbol{\pi} \in \boldsymbol{\Pi}_N^{HR}$ can be found which chooses slowscale actions as a function of the current state. Likewise, fastscale actions can be chosen as a function of the current state, the last slowscale action (and hence the state observed at the last slowscale epoch), and the time since the beginning of the renewal cycle. This, by definition, is a cyclo-stationary policy. □

Proof of Theorem 3.2 (DSM optimality equation). By Theorem 3.1, we can restrict the space of policies of the DSM from $\boldsymbol{\Pi}_N^{HR}$ to $\boldsymbol{\Pi}_N^{\mathrm{cyc}}$, without changing the optimal value. In this restricted model, $(N+1)$-step state transitions and rewards are as in a λ-discounted MDP with state space S, generic actions $\{a^\sigma, d_0, d_1, \dots, d_N\}$ and whose decision epochs are the renewal epochs.

The rewards and probabilities for the MDP are given by

$$\begin{aligned} &\tilde{r}(s, \{a^\sigma, d_0, d_1, \dots, d_N\}) = \\ &\qquad [r_{d_0} + P_{d_0} r_{d_1} + P_{d_0} P_{d_1} r_{d_2} + \dots P_{d_0} \cdots P_{d_{N-1}} r_{d_N}^{a^\sigma}](s) \end{aligned}$$

and

$$\tilde{p}(j | s, \{a^\sigma, d_0, d_1, \dots, d_N\}) = P_{d_0} P_{d_1} \cdots P_{d_{N-1}} P_{d_N}^{a^\sigma}(s, j)$$

The dynamic programming operator for this MDP is $\mathcal{L}_N$ and the optimal value is v_N^*. The assertions of the Theorem then follow from standard properties of discounted MDPs. □

Remark A.1 The proof of Theorem 3.2 identifies an equivalence between the discounted DSM and a discounted MDP with dynamic programming operator $\mathcal{L}_N$. From this equivalence, it is easy to see that any of the standard methods for solving discounted MDPs which make iterative use of dynamic programming operators – value iteration, policy iteration, etc ... – can be adapted to the DSM. For example, the value iteration sequence $x_n = \mathcal{L}_N^n x_0$ converges to v_N^*. Once v_N^* is known, an optimal cyclo-stationary policy can be found satisfying (3.5).

Remark A.2 From Theorem 3.2(b) or the proof of Theorem 3.1, it is evident that an optimal policy which prescribes the same a^σ in 2 different states can be modified so that it also prescribes the same sequences $\{d_0, d_1, \ldots, d_N\}$ in those states, with no loss in optimality. Since, ordinarily, $|A^\sigma| \leq |S|$, specifying an optimal policy therefore requires $O(N|A^\sigma||S|)$ data. In a DSM, therefore, the size of the policy space is much larger than that of the state space. It is magnified times N due to the time-inhomogeneity and times $|A^\sigma|$ due to the history dependent dynamics. Solution of the optimality problem can be expected to be correspondingly difficult computationally.

Remark A.3 While uniform optimality constitutes a stronger optimality criterion than conventional optimality, it is clear, in light of Theorem 3.3 and Remark A.1, that solving the conventional optimality problem (3.2), (3.3) tends to yield uniform optimal policies anyway. For in all standard algorithms, we first find v_N^* iteratively. To obtain an optimal cyclo-stationary policy we then evaluate the right hand side of (3.4) via dynamic programming. This generates sequences $\{a^\sigma, \pi, d\}$ satisfying (3.7) and (3.8).

Proof of Lemma 3.4 (Criteria for ϵ-optimality.)

(a) Equate the DSM to the standard discounted MDP described in the proof of Theorem 3.2. The result then follows from a standard construction of a stationary deterministic ϵ-optimal policy in discounted MDPs (see, for example, the proof of Theorem 6.2.11 in [8]).

(b) By the triangle inequality,

$$\begin{aligned} ||L^k L^{a^\sigma} v_N^* - L_\pi^k L_d^{a^\sigma} v_N^{\boldsymbol{\pi}_\epsilon^*}|| &\leq ||L^k L^{a^\sigma} v_N^* - L_\pi^k L_d^{a^\sigma} v_N^*|| + ||L_\pi^k L_d^{a^\sigma} v_N^* - L_\pi^k L_d^{a^\sigma} v_N^{\boldsymbol{\pi}_\epsilon^*}|| \\ &\leq (1-\lambda)\epsilon + \lambda ||v_N^* - v_N^{\boldsymbol{\pi}_\epsilon^*}||. \end{aligned}$$

The first term in the last inequality is due to the requirement of part (b). The second is due to the non-expansive property of dynamic programming operators.

Since the policy is ϵ-optimal, this becomes

$$||L^k L^{a^\sigma} v_N^* - L_\pi^k L_d^{a^\sigma} v_N^{\boldsymbol{\pi}_\epsilon^*}|| \leq \epsilon$$

which shows that it is uniform ϵ-optimal. $\square$

References

[1] E. Altman and V. Gaitsgory (1993) Control of a hybrid stochastic system. *Systems and Control Letters* **20** 307-314.

[2] E. Altman and V. Gaitsgory (1997) Asymptotic optimization of a nonlinear hybrid system governed by a Markov Decision Process. *SIAM J. Control Optim.* **35** 2070-2085.

[3] J.A. Filar, V. Gaitsgory, and A. Haurie (1997) Control of singularly perturbed hybrid stochastic systems. mimeo, submitted for publication.

[4] J.A. Filar and A. Haurie (1996) Optimal ergodic control of singularly perturbed hybrid stochastic systems. mimeo, submitted for publication.

[5] K. Hinderer and G. Hubner, On approximate and exact solution for finite stage dynamic programs, in *Markov Decision Theory*, edited by H. Tijms and J. Wessels, (Mathematical Centre, Amsterdam, 1977) pp. 57-76.

[6] 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

[7] M. Jacobson, N. Shimkin, A. Shwartz (1999) *Piecewise Stationary Markov Decision Processes, II.* Submitted for publication. Available at http://www.ee.technion.ac.il/~adam/PAPERS/2TimesScale2.ps

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

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

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

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

[12] P.J. Schweitzer (1988). Contraction mappings underlying undiscounted Markov decision problems – II. *J. Math. Anal. Appl.* **132** 154-170.

[13] J. Shapiro (1968). Turnpike planning horizons for a Markovian decision model *Man. Sci.* **14** 292-300.