MATHEMATICS OF OPERATIONS RESEARCH
Vol. 28, No. 4, November 2003, pp. 777–800
Printed in U.S.A.

MARKOV DECISION PROCESSES WITH SLOW SCALE PERIODIC DECISIONS

M. JACOBSON, N. SHIMKIN, AND A. SHWARTZ

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

For the PTMDP model, we consider the problem of optimizing the expected discounted reward when rewards devalue by a discount factor λ at the beginning of each slow scale cycle. When N is large, *initially stationary* policies (i.s.p.'s) are natural candidates for optimal policies. Similar to turnpike policies, an initially stationary policy uses the same decision rule for some large number of epochs in each slow scale cycle, followed by a relatively short planning horizon of time-varying decision rules. In this paper, we characterize the form of the optimal value as a function of N, establish conditions ensuring the existence of near-optimal i.s.p.'s, and characterize their structure. Our analysis deals separately with the cases where the time-homogeneous part of the system has state-dependent and state-independent optimal average reward. As we illustrate, the results in these two distinct cases are qualitatively different.

1. Introduction. Certain stochastic control problems involve the control of fast processes that are influenced by slower ones. The slow processes cause relatively infrequent perturbations in the usual operation of the faster processes. Such multiple time scale phenomena arise in a variety of ways in practice. The variety of models for such phenomena that have received attention in the literature is correspondingly rich. A number of authors (e.g., Davis 1993, Bäuerle 2001) have considered so-called Piecewise Deterministic Processes (PDPs) in which a deterministically controlled system is periodically perturbed by uncontrolled random events. Such models arise, for example, in manufacturing systems where the usual flow of production, a deterministic process, is occasionally perturbed due to random machine failure. Other authors have considered so-called hierarchical decision-making models (e.g., Sethi and Zhang 1994) in which the control space consists of several different types of decision variables used to control related processes. Typically, the different processes are controlled by decisions made with different frequency, leading to multiple time scales. These models suit, for example, manufacturing systems where the actual production operations constitute a process run by frequent, short-run decisions while the resources available for production (e.g., number of machines, workers, etc.) are in the control of less frequent, long-run decisions.

In some literature, multiple time scale phenomena have been treated in a Markov Decision Process (MDP) framework. Singularly perturbed MDPs (e.g., Delebecque and Quadrat 1981, Phillips and Kokotovic 1981) are perhaps the most prominent example, and have been used to model applications such as the control of queuing systems and hydroelectric power

Received November 20, 1999; revised April 18, 2002, and March 3, 2003.
MSC2000 subject classification. Primary: 60J05.
OR/MS subject classification. Primary: Probability/Markov processes.
Key words. Periodic time-inhomogeneity, multiple time scales, turnpike, cyclo-stationary, discounted cost, multi-class models.

0364-765X/03/2804/0777
1526-5471 electronic ISSN,

generation operations. In these models, slow scale phenomena correspond to infrequent, but time-homogeneous transitions made between certain disjoint subsets of the state space.

In this paper, we consider a discrete-time, two time scale MDP model in which slow scale phenomena are modeled by time-inhomogeneity in the state transition and reward data. Namely, the decision-making horizon can be partitioned into intervals of $N+1$ epochs called *slow scale cycles*. Over the first N epochs of each slow scale cycle, decisions of a fast scale type are made and their effect is described by time-homogeneous state transition and reward functions. At the final epoch, decisions of a distinct, slow scale type are made and described by different transition and reward functions. The process repeats in an $(N+1)$-periodic fashion. We call these models Periodically Time-Inhomogeneous Markov Decision Processes (PTMDPs), so as to reflect that the decision process is time-homogeneous except for the periodic occurrence of the slow scale decision epochs. A PTMDP is essentially a Markovian, two-level hierarchical decision-making model in discrete time. The model assumes that the times between slow-scale epochs are all equal, although generalizations are possible.

For the PTMDP model, we address the problem of optimizing the expected discounted reward when rewards devalue by a discount factor λ at the beginning of each slow scale cycle. This criterion when the slow scale epochs demarcate time periods over which total performance is of interest and when high performance in early periods is a priority. Between slow epochs, rewards are time-homogeneous (with no discounting). We emphasize that, although the length of the slow period may be regarded as a parameter, the discount factor is held fixed throughout. A variation of this criterion is also briefly discussed (see Remark 5). Our aim is to understand how the solutions of the problem behave for large, finite values of N, i.e., when the interval of fast scale epochs in each slow scale cycle is long. We assume finite state and action spaces throughout.

As a motivating example, we consider the application of this model to what we call multiple project management problems. In these problems, the decision maker controls one of m different projects for a period of N epochs, at the end of which (s)he may select a different project. In this case, the slow scale decisions are the different project selections. The fast scale decisions are the short-run decisions controlling the current project. Our discounted optimality criterion is apt, for example, when the management team is subject to an evaluation at the conclusion of each project. It is then natural to look at the net outcome of the successive projects. This discounting reflects that a favorable evaluation depends more critically on success in earlier projects than later ones.

As usual, we are interested both in the optimal value obtainable and the decision-making policies that attain it. Due to the $(N+1)$-periodic nature of the problem, one can expect, at the outset, that $(N+1)$-periodic sequences of decision rules will be the simplest kind of optimal policy in this model. On the other hand, a PTMDP is, crudely speaking, a sequence of N-horizon time-homogeneous MDPs, terminated by an epoch with different reward and transition data. Intuition therefore suggests that for large N, a policy of a simplified turnpike-like form, which we call an *initially stationary policy* (i.s.p.), will result in near optimality in the discounted PTMDP. By an i.s.p., we mean a policy which, in each slow scale cycle, uses a time-independent decision rule for some large number of epochs followed by a relatively short planning horizon of time-dependent decision rules.

The notion of turnpikes originates in the work of Shapiro (1968), and was subsequently expanded by Hinderer and Hubner (1977). 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-independent 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-dependent decision rules may be required. This intuition was verified in Hinderer and Hubner (1977) and Shapiro (1968) for discounted MDPs possessing a unique

optimal decision rule. The time-independent decision rule is referred to as a "turnpike," while the interval of time-dependent decision rules is referred to as the *planning horizon*. Analogous results hold for nondiscounted MDPs as well.

The same intuition leads us to conjecture that i.s.p.'s may be nearly optimal in PTMDPs with large N. For in these models, the decision maker faces a long, time-homogeneous horizon at the beginning of each slow scale cycle. In this paper, we investigate conditions ensuring the existence of (approximately) optimal i.s.p.'s. One can anticipate that the existence of approximately optimal i.s.p.'s will bring all the usual advantages of a simplified policy structure. Namely, the policy is easier to implement and simplified algorithms may be used to compute it.

A history-dependent variant of the model of the present paper has been considered in Jacobson (1998). In that model, the slow scale decisions take effect N epochs after they are made. The main results for this model are similar to the present ones.

The rest of this paper is organized as follows. In §2, we present the PTMDP model and describe a generic multiple project management problem to motivate it. Section 3 introduces notation and conventions used throughout. Section 4 describes our discounted optimality criterion and summarizes certain analogies between optimal solutions in PTMDPs and those in standard discounted MDPs. Sections 5 and 6 contain our analysis, which is carried out under certain conditions on Ψ, the average reward MDP associated with the fast scale data. No assumptions are made about the slow scale data. The analysis is divided into two cases. The first case is when Ψ has an optimal gain vector that is independent of the initial state. The second case is when the optimal gain may be state-dependent.

Section 5 addresses the state-independent gain case. We begin by characterizing the form of the optimal value as a function of N. This structure is used to establish the existence of ϵ-optimal initially stationary policies with a finite planning horizon which depends 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. 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 that are maximizing in Ψ's optimality equation is a singleton.

Section 6 addresses the state-dependent gain case. The analysis is carried out under less general conditions than the state-independent case, however, these conditions are satisfied by certain applications of interest such as multiple project management. We again analyze the form of the optimal value as a function of N. In addition, we establish that there exists a uniformly $N\epsilon$-optimal i.s.p. whose planning horizon, η_ϵ, is a function only of ϵ. Moreover, the turnpike decision rule may be derived from the fast scale MDP, Ψ, alone. The reader will note that the i.s.p. is only $N\epsilon$-optimal rather than ϵ-optimal. However, this kind of scaled ϵ-optimal is still reasonable since the optimal value vector also scales with N.

In §7, we briefly discuss conditions for the existence of uniformly ϵ-optimal initially stationary policies. This subject has been more fully explored in Jacobson (1998). Section 8 examines the degree of suboptimality of a policy that uses the same decision rule at all fast scale epochs. Such a policy is the closest analogue to a stationary policy from classical MDP theory. It is therefore natural to wonder how suboptimal these policies might be in a "nearly time-homogeneous" decision process like a PTMDP. Finally, in §9, we summarize approaches to computing approximately optimal initially stationary policies.

2. Model introduction. A Periodically Time-Inhomogeneous Markov Decision Process (PTMDP) is a discrete time MDP whose evolution is described by an $(N+1)$-periodic sequence of transition probability and reward functions. Over the first N epochs (called *fast scale epochs*) of each periodic interval, the probability and reward functions are time-homogeneous. At the final epoch, however, they are distinct. We call these distinct epochs

slow scale epochs. In addition, we refer to the $(N+1)$-periodic intervals as *slow scale cycles*.

Formally, the process is defined by a state space S, and for each $s \in S$ a space of fast scale actions A_s and slow scale actions A_s^σ. At time t and state $s_t \in S$, a fast scale action $a_t \in A_{s_t}$ is selected if t is a fast scale epoch. The rewards and transition probabilities are given by fast scale data $r(s_t, a_t)$, $p(s_{t+1}|s_t, a_t)$. Conversely, if t is a slow scale epoch, a slow scale action $a_t^\sigma \in A_{s_t}^\sigma$ is selected and the rewards and transition probabilities are given by slow scale data $r^\sigma(s_t, a_t^\sigma)$, $p^\sigma(s_{t+1}|s_t, a_t^\sigma)$. We shall assume throughout that all state and action spaces are finite. Hence, the reward data has a bound, which we denote $r_{\max}$. The PTMDP timing framework is depicted for one slow scale cycle in Figure 1.

EXAMPLE 1 (MULTIPLE PROJECT MANAGEMENT). As a motivating example of a PTMDP, we describe what we call a *multiple project management problem*. In problems of this type, work is done on one of m different projects for a duration of N decision epochs. At epoch $N+1$, a transition to a different project may take place, and the process repeats.

Formally, each project is modeled by one of m MDPs, $\{\Psi^i\}_{i=1}^m$. These MDPs have associated state and action spaces $\{S^i, A_s^i\}_{i=1}^m$ and associated reward and probability functions $\{r^i(s,a), p^i(j|s,a)\}_{i=1}^m$. The multiple project management problem can then be modeled as a PTMDP in which the fast scale epochs are those at which work is done on one of the projects, while the slow scale epochs are those at which transitions between projects may occur. The state and fast scale action spaces are

$$S = \bigcup_{i=1}^m S^i$$
$$A_s = A_s^i, \qquad s \in S^i$$

and the fast scale reward and probability functions are

$$r(s,a) = r^i(s,a), \qquad s \in S^i$$
$$p(j|s,a) = \begin{cases} p^i(j|s,a) & s, j \in S^i \\ 0 & s \in S^i, \quad j \notin S^i \end{cases}$$

We make no specific requirements on A^σ, $r^\sigma(s, a^\sigma)$, or $p^\sigma(j|s, a^\sigma)$. Typically, however, $p^\sigma(j|s, a^\sigma)$ would allow transitions between different S^i, corresponding to changes from one project to another. The slow scale rewards $r^\sigma(s, a^\sigma)$ might specify startup costs for a new project.

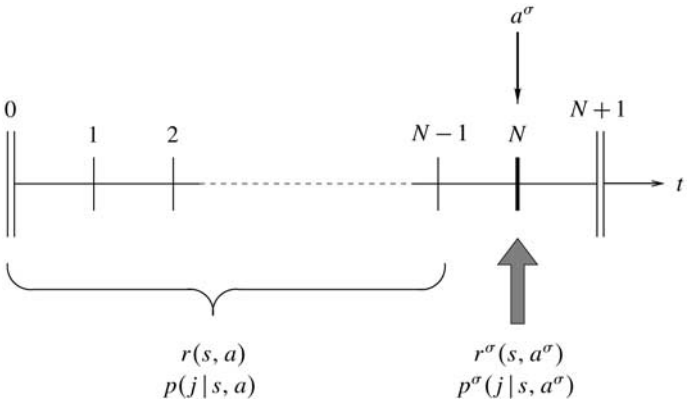

FIGURE 1. Time evolution of a PTMDP.

3. Notation and preliminaries.

3.1. Norms and dynamic programming notation. We shall represent functions on the state space 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 $\mathbb{R}^{|S|}$. For a vector $y \in \mathbb{R}^{|S|}$, the max-norm and span semi-norm are

$$\|y\| \stackrel{\triangle}{=} \max_{s\in S} |y(s)|$$

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

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 $\bar{\nu}(c)$ to denote an unspecified vector in $\mathbb{R}^{|S|}$ whose max-norm is at most c.

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

$$Lx(s) \stackrel{\triangle}{=} \max_{a\in A_s} \left\{ r(s,a) + \sum_{j\in S} p(j|s,a)x(j) \right\}, \qquad s \in S. \tag{3.1}$$

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

Property (ii) is sometimes referred to as the *nonexpansive property* of the L operator. From this property, we deduce that for any vector y and any vector x with $\|x\| \leq c$, $L(y + x) = Ly + \bar{\nu}(c)$.

For a given $0 \leq \lambda < 1$, we define the discounted dynamic programming operator for the slow scale data

$$L^\sigma x(s) \stackrel{\triangle}{=} \max_{a^\sigma \in A_s^\sigma} \left\{ r^\sigma(s,a^\sigma) + \lambda \sum_{j\in S} p^\sigma(j|s,a^\sigma)x(j) \right\}, \qquad s \in S. \tag{3.2}$$

Finally, we define the $(N+1)$-step PTMDP discounted operator $\mathscr{L}_N$,

$$\mathscr{L}_N x \stackrel{\triangle}{=} L^N L^\sigma x. \tag{3.3}$$

It corresponds to backward induction over one slow scale cycle.

3.2. Policies and decision rules. A history-dependent, randomized policy shall be defined, as usual, as a sequence of maps which specify, for each possible state-action history, a probability distribution over the relevant action set. This set is A_{s_t} when t is a fast scale epoch, and $A_{s_t}^\sigma$ otherwise. The space of history-dependent, randomized policies for a PTMDP whose slow scale cycles are of length $N+1$ shall be denoted $\mathbf{\Pi}_N^{HR}$.

A Markovian deterministic decision rule, d, 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. (Henceforth, the term *decision rule* shall implicitly refer to a Markovian deterministic one, unless otherwise stated.) Let D denote the space of fast scale decision rules mapping each $s \in S$ to some $a \in A_s$. Similarly, let D^σ denote the space of slow scale 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 a short-hand for (3.1). Similar notation will be used for the other dynamic programming operators. Note that $Lx \geq L_d x$ with equality for at least one d. A decision rule, $d \in D$, attaining this maximum shall be called an x-improving or x-maximizing decision rule.

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

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

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.

For a PTMDP with slow scale cycle length N, we define a cyclo-stationary policy to be a policy that uses a common $(N+1)$-length decision rule sequence $\{d_0, d_1, \ldots, d_N, d^\sigma\}$ in each slow scale cycle. The set of all cyclo-stationary policies shall be denoted $\mathbf{\Pi}_N^{\text{cyc}}$. A cyclo-stationary policy shall be abbreviated as $\{\pi, d^\sigma\}$ where $\pi = \{d_0, d_1, \ldots, d_{N-1}\}$, $d_i \in D$ are the decision rules used at the fast scale epochs while $d^\sigma \in D^*$ is the decision rule used at the slow scale epoch.

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

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

That is, the policy prescribes some time-invariant decision rule, δ, for all but possibly the final η fast scale epochs ($\eta < N$) of the slow scale cycle. We refer to η as the i.s.p.'s *planning horizon*. 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. (borrowing from turnpike terminology). Also, we refer to δ as the *turnpike decision rule*.

3.3. The fast scale MDP. The fast scale data, $r(s_t, a_t)$, $p(s_{t+1}|s_t, a_t)$, in a PTMDP can be associated with an average reward MDP, denoted as Ψ, with decision rule space D. We shall refer to Ψ as the *fast scale* MDP of the PTMDP.

The gain vector g_d of a decision rule $d \in D$ is defined as $g_d \stackrel{\triangle}{=} P_d^* r_d$ where $P_d^* = \lim_{n\to\infty} (1/n) \sum_{k=0}^n P_d^k$ is the limiting matrix of P_d. Obviously, $g_d(s)$ is the average reward from initial state s. Since the action and state spaces are finite, it is well known that g^*, the optimal gain vector of Ψ, satisfies $g^* \geq g_d$ with equality for some decision rule d. We define $D^* \stackrel{\triangle}{=} \{d \in D : g^* = g_d\}$ as the set of optimal decision rules in Ψ.

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

$$g^* = Ug^* \stackrel{\triangle}{=} \max_{d \in D} \{P_d g^*\} \tag{3.4}$$

$$g^* + v = Tv \stackrel{\triangle}{=} \max_{d \in E} \{r_d + P_d v\} \tag{3.5}$$

Here, $E \stackrel{\triangle}{=} \{d \in D : P_d g^* = g^*\}$ and the operators U and T are defined as indicated. The set

$$V \stackrel{\triangle}{=} \left\{v \in \mathbb{R}^{|S|} : g^* + v = Tv\right\}$$

of solutions to (3.5) is a closed, unbounded Schweitzer and Federgruen (1978) set. For a given $v \in V$, we denote

$$E(v) \stackrel{\Delta}{=} \{d \in E : g^* + v = T_d v\}.$$

It is a standard result that $D^* \subset E$ and also that $E(v) \subset D^*$ for any such $v \in V$. Hence, any maximizing decision rule in (3.5) is average optimal.

Define

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

In Schweitzer and Federgruen (1978), 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 in Schweitzer and Federgruen 1978):

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 [Schweitzer and Federgruen 1978, p. 314]). Based on that paper, it is also known 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 Schweitzer-Federgruen class, $R^*(\alpha)$. Furthermore, if there is only a single Schweitzer-Federgruen class (i.e., $n^* = 1$), then given any $v_0 \in V$, $V = \{v \in \mathbb{R}^{|S|} : v = v_0 + c\mathbf{1}, c \in \mathbb{R}\}$. The converse is also true, namely, V is one-dimensional only if $n^* = 1$.

3.4. Zero-reward analogues. When we consider a version, Ψ_0, of Ψ in which all rewards are set to zero, we obtain useful analogues to the properties cited in Section 3.3. Firstly, equations (3.4) and (3.5) reduce to

$$w = \max_{d \in D} \{P_d w\} = Uw \tag{3.6}$$

where w has replaced v. The set $W \stackrel{\Delta}{=} \{w \in R^{|S|} : w = Uw\}$ is the analogue of V. Observe that $\mathbf{1} \in W$. Also, from to (3.4), one can see that $g^* \in W$. (Note that g^* still refers to the optimal gain of Ψ, not Ψ_0.)

Since U is *positive homogeneous*, i.e., $Ucx = cUx$ for any nonnegative scalar c, it follows that W is a cone. For any $w \in W$, define

$$K(w) \stackrel{\Delta}{=} \{d \in D : P_d w = Uw = w\}$$

as the set of decision rules attaining the maximum in (3.6).

Since all decision rules are optimal in Ψ_0, we have the following analogue of the set R^*,

$$\hat{R} \stackrel{\Delta}{=} \{s \in S : s \text{ is recurrent for some decision rule}\}.$$

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 that 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, Schweitzer 1984).

By analogy with V, the difference $w_1(s) - w_2(s)$ for any $w_1, w_2 \in W$ is independent of s over each fixed Bather class, $\hat{R}(\alpha)$. Since $\mathbf{1} \in W$, it follows that if $w \in W$ then $w(s)$ is constant as a function of s on each $\hat{R}(\alpha)$. Furthermore, when there is only a single Bather class (i.e., $\alpha^* = 1$) then $W = \{w \in \mathbb{R}^{|S|} : w = c\mathbf{1}, c \in \mathbb{R}\}$. A stationary MDP for which $\alpha^* = 1$ is called *weakly communicating*.

REMARK 1. From property (1), it follows that given any $\alpha_0 \in \{1, \ldots, n^*\}$, then $R^*(\alpha_0) \subset \hat{R}(\alpha)$ for some $\alpha \in \{1, \ldots, \alpha^*\}$.

3.5. Assumptions. We shall work with combinations of the following conditions on the fast scale MDP, Ψ. See §§3.3 and 3.4 for relevant notation and concepts.

CONDITION 1. *The optimal gain vector g^* has the form $g_0\mathbf{1}$, i.e., it is state-independent.*

CONDITION 2. *V is one dimensional, i.e., for some v_0, $V = \{v \in \mathbb{R}^{|S|} : v = v_0 + c\mathbf{1}, c \in \mathbb{R}\}$.*

CONDITION 3. *The sequence $L^k x - kg^*$ converges, as k tends to infinity, for all vectors x.*

CONDITION 4. *The sequence $U^k x$ converges, as k tends to infinity, for all vectors x.*

CONDITION 5. *There exists an average optimal decision rule, $\gamma \in D^*$, satisfying, for all $w \in W$,*

$$P_\gamma w = w = Uw. \tag{3.7}$$

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

Perhaps the most general, commonly considered class of models in which Condition 1 holds is the class of weakly communicating models (see Puterman 1994, pp. 348–352). The MDP Ψ is weakly communicating if it has a single Bather class (i.e., $\alpha^* = 1$).

Condition 2 is also sufficient for Condition 1. As mentioned in §3.3, Condition 2 holds if and only if there is a single Schweitzer-Federgruen class (i.e., $n^* = 1$). If $n^* = 1$, then a randomized, unichain gain optimal decision rule exists, implying Condition 1. A commonly considered condition which implies Condition 2 is when Ψ is a unichain model, meaning that P_d has a single recurrent class for all $d \in D$.

Conditions 3 and 4 are equivalent to nonperiodicity requirements on the chain structure of Ψ. Condition 3 is equivalent [10] to the condition that a randomized, aperiodic, gain optimal decision rule exists whose recurrent chains are the Schweitzer-Federgruen classes $R^*(\alpha)$, $\alpha = 1, \ldots, n^*$. By analogy with an MDP with zero rewards, Condition 4 is equivalent to the condition that a randomized, aperiodic decision rule (not necessarily gain optimal) exists whose subchains are $\hat{R}(\alpha)$, $\alpha = 1, \ldots, \alpha^*$. A simple condition for which Conditions 3 and 4 will hold is if P_d is aperiodic for all $d \in D$.

Condition 5 is one which, to our knowledge, has not been considered before. The following two hypotheses together imply Condition 5.

HYPOTHESIS 1. *All states are recurrent for at least one decision rule, i.e., $S = \hat{R}$.*

HYPOTHESIS 2. *There is a Schweitzer-Federgruen class in each Bather class, i.e., $\hat{R}(\alpha) \cap R^* \neq \{\varnothing\}$, for every $\alpha = 1, \ldots, \alpha^*$.*

When Hypothesis 1 holds, the Bather classes completely partition the state space, S. When Hypothesis 2 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 (3.7).

REMARK 2. Hypothesis 2 holds trivially when Ψ is weakly communicating. 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 $d \in D^*$ must satisfy $P_d^* g^* = g^*$.

Another hypothesis that implies Condition 5 is

HYPOTHESIS 3. *The state space has a partition into m subchains, $\{S^i\}_{i=1}^m$, each of which contains a single Bather class and is globally closed, i.e., the S^i do not communicate with one another under any decision rule.*

When Hypothesis 3 holds, Ψ can be viewed as a disjoint collection of weakly communicating sub-MDPs, Ψ^i, with state spaces S^i. As mentioned in §3.4, the set W in a weakly communicating MDP is identically the set of state-independent vectors in $\mathbb{R}^{|S|}$. For MDPs satisfying Hypothesis 3, this property has the following obvious generalization,

$$W = \left\{ w \in \mathbb{R}^{|S|} : \exists\, c^i \in \mathbb{R}, i = 1, \ldots, m \text{ s.t. } w(s) = c^i \text{ if } s \in S^i \right\}$$

Therefore, for any $w \in W$, all $\gamma \in D$ (and not just optimal ones) satisfy (3.7), so that Condition 5 follows.

In a multiple project management problem (see Example 1), Hypothesis 3 is satisfied if the MDPs Ψ^i are weakly communicating. In this case, the sets S^i are the project state spaces.

4. Optimality criteria. In this section, we describe a discounted optimality criterion for the PTMDP model. Propositions are presented that follow by analogy with standard discounted MDP theory, and therefore we omit proofs.

Consider a PTMDP and let $0 \leq \lambda < 1$ be a discount factor. For every history-dependent, randomized policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}_N^{HR}$, we define the following value function

$$v_N^{\boldsymbol{\pi}}(s) \stackrel{\triangle}{=} E_s^{\boldsymbol{\pi}} \left(\sum_{t=0}^{\infty} \lambda^{\lfloor t/(N+1) \rfloor} \boldsymbol{r}_t \right), \qquad s \in S\,, \tag{4.1}$$

where $E_s^{\boldsymbol{\pi}}$ is the expectation operator induced by $\boldsymbol{\pi}$ starting in state s, and $\boldsymbol{r}_t$ is the associated stochastic process of rewards.

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

$$v_N^*(s) \stackrel{\triangle}{=} \sup_{\boldsymbol{\pi} \in \boldsymbol{\Pi}_N^{HR}} v_N^{\boldsymbol{\pi}}(s) \tag{4.2}$$

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

$$v_N^*(s) - v_{\boldsymbol{\pi}_{\boldsymbol{\epsilon}}^*}(s) \leq \boldsymbol{\epsilon} \tag{4.3}$$

for a given $\boldsymbol{\epsilon} \geq 0$. Policies $\boldsymbol{\pi}_{\boldsymbol{\epsilon}}^*$ satisfying (4.3) are said to be $\boldsymbol{\epsilon}$-optimal. A 0-optimal policy is said to be optimal.

Equations (4.1), (4.2), and (4.3) define a discounted reward optimality criterion in which the rewards devalue by λ at the start of each slow scale cycle. This criterion is appropriate when the slow scale epochs demarcate time periods over which total performance is of interest, and when high performance in early periods is a priority. As an example application, one may consider a multiple project management problem (see Example 1) in which projects are run for a year (a period subdivided into N fast scale decision epochs) and subject to annual evaluation. The criterion reflects that a high total performance from projects in earlier years, rather than later ones, leads to a more favorable evaluation.

The following proposition characterizes solutions of the optimality problem.

PROPOSITION 4.1 (PTMDP OPTIMALITY EQUATION).

(a) *The operator $\mathscr{L}_N$ is a contraction operator with respect to the max norm $\|\cdot\|$, with contraction factor λ and fixed point v_N^* (the optimal value defined in* (4.2)*). Hence,*

$$v_N^* = \mathscr{L}_N v_N^* = L^N L^\sigma v_N^*. \tag{4.4}$$

(b) *A cyclo-stationary policy $\{\pi, d^\sigma\}$ is optimal if and only if*

$$v_N^* = L_\pi^N L_{d^\sigma}^\sigma v_N^*. \tag{4.5}$$

That is, $\{\pi, d^\sigma\}$ is conserving.

Clearly, the maximizations implicit in the right hand equality of (4.4) are attained by some cyclo-stationary policy $\{\pi, d^\sigma\}$, due to the finiteness of the state and action spaces. Hence, by part (b) of the Proposition, it follows that at least one optimal cyclo-stationary policy always exists.

REMARK 3. In classical discounted MDP theory, algorithms for computing the optimal value and policy are driven by iterative operations involving the model's discounted dynamic programming operator. Analogous algorithms are available for solving PTMDPs that manipulate $\mathscr{L}_N$ in parallel ways. For example, the analogue of value iteration would be to compute the sequence $\mathscr{L}_N^n x_0$ for some initial vector x_0. The sequence converges to v_N^* as n goes to infinity.

A cyclo-stationary policy $\{\pi, d^\sigma\}$ shall be called *uniformly* ϵ-optimal if it is ϵ-optimal starting from *any* time epoch (and not just the initial one). This is the case when (4.3) holds, namely $\|v_N^* - v_N^{\pi_\epsilon^*}\| \le \epsilon$, and further

$$\left\| L^k L^\sigma v_N^* - L_\pi^k L_{d^\sigma}^\sigma v_N^{\pi_\epsilon^*} \right\| \le \epsilon \tag{4.6}$$

for all $0 \le k < N$.

The following proposition provides sufficient conditions, in terms of v_N^*, for cyclo-stationary policies to be ϵ-optimal and uniformly ϵ-optimal.

PROPOSITION 4.2 (CRITERIA FOR ϵ-OPTIMALITY).

(a) *Suppose that a cyclo-stationary policy $\{\pi, d^\sigma\}$ satisfies*

$$\|v_N^* - L_\pi^N L_{d^\sigma}^\sigma v_N^*\| \le (1-\lambda)\epsilon.$$

Then the cyclo-stationary policy is ϵ-optimal.

(b) *Suppose, in addition, that*

$$\|L^k L^\sigma v_N^* - L_\pi^k L_{d^\sigma}^\sigma v_N^*\| \le (1-\lambda)\epsilon$$

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

REMARK 4. Proposition 4.2(b) indicates that, when v_N^* is known, uniformly ϵ-optimal policies can be obtained by carrying out the maximizations implicit in the expression $L^k L^\sigma v_N^*$ up to an accuracy of $(1-\lambda)\epsilon$. This means that value iteration strategies—in which one first finds an approximation to v_N^* and then extracts from it a policy via dynamic programming—will generally produce uniformly ϵ-optimal policies.

Such a strategy can fail to find ϵ-optimal policies having a particularly simple form, but which are not uniformly optimal (see, for example, Jacobson 1998, Example 8.18). However, policies of this latter type are of questionable usefulness, because they do not prescribe optimal decisions from every starting state and time. In practice, one would need to discard them if certain unlikely state transitions are made.

REMARK 5. A variant of (4.1) that may be natural to consider is

$$v_N^{\boldsymbol{\pi}}(s) \stackrel{\triangle}{=} E_s^{\boldsymbol{\pi}} \sum_{t=0}^{\infty} \lambda^{t/(N+1)} \boldsymbol{r}_t, \qquad s \in S\,. \tag{4.7}$$

In this case, the discount factor applied at the end of each slow scale cycle is still λ, irrespective of N. However, the discounting now occurs gradually over all fast scale epochs at rate $\beta_N = \lambda^{1/(N+1)}$.

For this variant, the optimality equation still has the form $v_N^* = L^N L^\sigma v_N^*$, but with the revised definitions

$$Lx(s) \stackrel{\triangle}{=} \max_{a \in A_s} \left\{ r(s,a) + \beta_N \sum_{j \in S} p(j|s,a) x(j) \right\} \tag{4.8}$$

$$L^\sigma x(s) \stackrel{\triangle}{=} \max_{a^\sigma \in A_s^\sigma} \left\{ r^\sigma(s,a^\sigma) + \beta_N \sum_{j \in S} p^\sigma(j|s,a^\sigma) x(j) \right\}. \tag{4.9}$$

These operators are more complicated than those defined in (3.1) and (3.2) in that they now depend on N through the discount factor β_N. Since $\beta_N \to 1$ as $N \to \infty$, this does not fall into the standard discounted MDP formulation. In particular, the convenient "tail effect" of standard discounted costs is not present on the fast time scale (within a single slow scale cycle). The analysis of the modified criterion is therefore not simpler than that of criterion (4.1) (to follow in §§5 and 6), but rather requires additional consideration for the variation in β_N. On heuristic grounds, we conjecture that similar results hold regarding the existence of i.s.p.'s.

In the context of specific applications, the specific choice of discounting within a slow-scale cycle may often be a modeling choice rather than a dictate of the application.

5. Analysis of the state-independent gain case. In this section, we address PTMDPs for which the optimal gain of the fast scale MDP is state-independent, i.e., for which Condition 1 holds. In addition, we shall also assume Condition 3 throughout. However, Condition 3 is largely a technical condition that removes periodicity effects from consideration. When Condition 3 does not hold, generalizations to account for periodicity are possible, although trite.

Since Condition 1 holds, then the set E in (3.5) coincides with D, so that (3.4) and (3.5) reduce to

$$g^* + v = Lv.$$

Hence, $E(v)$ reduces to the set of decision rules which are v-improving with respect to the L operator. For any $v \in V$, we shall define

$$\Delta(v) \stackrel{\triangle}{=} \frac{1}{2}\Big(\min_{d \in D \setminus E(v)} \|Lv - L_d v\| \Big) \tag{5.1}$$

with $\Delta(v) = \infty$ if $E(v) = D$. Observe that

$$\arg\max_{d \in D} \{r_d + P_d v + \bar{\nu}(\Delta(v))\} \subset E(v). \tag{5.2}$$

Equations like (5.2) are key to many turnpike results in the literature. We shall use it in a similar way to establish Theorem 5.3.

Since Condition 3 holds, the operator

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

is defined on all $\mathbb{R}^{|S|}$. As established in Schweitzer and Federgruen (1979), convergence to $\hat{L}^\infty x$ is geometric with a rate that depends on x. Furthermore, it is known (cf. Lemma 2.2(g) in Schweitzer and Federgruen 1979) that $\hat{L}^\infty x$ is a mapping from $\mathbb{R}^{|S|}$ to V and (cf. Lemma 2.1 in Schweitzer and Federgruen 1979) that $\hat{L}^\infty$ has the same properties as L (in particular, properties (i) and (ii) in §3.1).

We now define y_∞ as the unique vector satisfying the fixed point equations

$$y_\infty = \hat{L}^\infty L^\sigma y_\infty . \tag{5.3}$$

The uniqueness of y_∞ follows by noting that the operator $\hat{L}^\infty L^\sigma$ is a contraction operator with respect to $\|\cdot\|$ with contraction factor λ. It is clear from (5.3) that $y_\infty \in V$ since, as mentioned above, $\hat{L}^\infty$ maps into V. We also define

$$\begin{aligned} M_{y_\infty}(\epsilon) &\stackrel{\triangle}{=} \min\{k : \|L^k L^\sigma y_\infty - kg^* - \hat{L}^\infty L^\sigma y_\infty\| \le \epsilon\} \\ &= \min\{k : \|L^k L^\sigma y_\infty - kg^* - y_\infty\| \le \epsilon\}. \end{aligned}$$

By Lemma 2.2(e) in Schweitzer and Federgruen (1979), if Condition 1 holds, the sequence $\|L^k x - kg^* - \hat{L}^\infty x\|$ decreases monotonically in k. Hence, $M_{y_\infty}(\epsilon)$ is the smallest k for which $L^k L^\sigma y_\infty - kg^*$ has converged within ϵ.

The following theorem describes the asymptotic behavior of the optimal value vector v_N^*. It establishes that $v_N^* - Ng^*/(1-\lambda)$ converges to y_∞ geometrically.

THEOREM 5.1 (ASYMPTOTIC BEHAVIOR OF v_N^*, STATE-DEPENDENT GAIN). *Assume Conditions* 1 *and* 3 *hold, fix* $\epsilon > 0$, *and suppose that* $N \ge M_{y_\infty}(\epsilon)$. *Then*

$$v_N^* = \frac{Ng^*}{1-\lambda} + y_\infty + \bar{\nu}\left(\frac{\epsilon}{1-\lambda}\right). \tag{5.4}$$

PROOF. Let $z_0 = Ng^*/(1-\lambda) + y_\infty$. Hence,

$$\begin{aligned} \mathcal{L}_N z_0 &= L^N L^\sigma z_0 \\ &= L^N L^\sigma y_\infty + \frac{\lambda N g^*}{1-\lambda} && (5.5) \\ &= \hat{L}^\infty L^\sigma y_\infty + Ng^* + \bar{\nu}(\epsilon) + \frac{\lambda N g^*}{1-\lambda} && (5.6) \\ &= z_0 + \bar{\nu}(\epsilon). && (5.7) \end{aligned}$$

In the above, (5.5) followed from Condition 1, (5.6) followed from our assumption that $N \ge M_{y_\infty}(\epsilon)$, and (5.7) followed from (5.3) and rearranging terms. By Proposition 4.1, the Banach theorem then implies,

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

from which (5.4) follows. □

REMARK 6. The Theorem implies that for any $\epsilon > 0$

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

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

Our next result, Theorem 5.2, indicates that an ϵ-optimal i.s.p. exists whose planning horizon η_ϵ is bounded independently of N. In addition, the turnpike decision rules are gain optimal for Ψ.

When Condition 3 holds, the right hand side of (4.4) gives us some reason to anticipate such a result mathematically. If the sequence $L^k L^\sigma v_N^* - kg^*$ has nearly converged after $\eta < N$ steps, then $L^\eta L^\sigma v_N^*$ approximates an element of V, and a $\delta \in D^*$ exists satisfying

$$L_\delta^{N-\eta} L^\eta L^\sigma v_N^* \approx L^N L^\sigma v_N^*. \tag{5.8}$$

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

For (5.8) to be satisfied, we might initially expect to require some condition which bounds the rate of convergence of the sequence $L^k L^\sigma v_N^* - kg^*$. Otherwise, since v_N^* is a priori unknown, we cannot be sure that convergence will take place in $\eta < N$ steps. A virtue of the analysis that follows is that it does not require any such conditions. Apart from the state-independence of g^*, Theorem 5.2 requires only Condition 3, the existence of the limit.

THEOREM 5.2 (EXISTENCE OF ϵ-OPTIMAL I.S.P.'S). *Assume that Conditions* 1 *and* 3 *hold. Fix* $\epsilon > 0$, $\delta \in E(y_\infty)$, *and let* $\eta_\epsilon = M_{y_\infty}((1-\lambda)^2\epsilon/2)$. *Then if* $N > \eta_\epsilon$, *a uniformly* ϵ*-optimal i.s.p. exists with planning horizon* η_ϵ *and turnpike decision rule* δ.

PROOF. Fix any $\delta \in E(y_\infty)$ and let $d^\sigma \in D^\sigma$ satisfy $L^\sigma y_\infty = L_{d^\sigma}^\sigma y_\infty$. Let π be a sequence of N decision rules whose first $N - \eta_\epsilon$ terms are δ, and which satisfies $L^k L^\sigma y_\infty = L_\pi^k L_{d^\sigma}^\sigma y_\infty$ for $0 \leq k \leq \eta_\epsilon$. Since $N > \eta_\epsilon = M_{y_\infty}((1-\lambda)^2\epsilon/2)$, we have from Theorem 5.1 that

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

In light of Condition 1, the last equation implies,

$$\|L^k L^\sigma v_N^* - L_\pi^k L_{d^\sigma}^\sigma v_N^*\| \leq \|L^k L^\sigma y_\infty - L_\pi^k L_{d^\sigma}^\sigma y_\infty\| + \lambda(1-\lambda)\epsilon, \tag{5.9}$$

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 (5.9) is zero. Therefore, we deduce that

$$\|L^k L^\sigma v_N^* - L_\pi^k L_{d^\sigma}^\sigma v_N^*\| \leq (1-\lambda)\epsilon. \tag{5.10}$$

Furthermore, by our definitions of η_ϵ and y_∞, we have that, for $k \geq \eta_\epsilon$,

$$\begin{aligned} L^k L^\sigma y_\infty &= \hat{L}^\infty L^\sigma y_\infty + kg^* + \bar{\nu}\left((1-\lambda)^2\epsilon/2\right) \\ &= y_\infty + kg^* + \bar{\nu}\left((1-\lambda)^2\epsilon/2\right) \end{aligned} \tag{5.11}$$

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

$$L^{\eta_\epsilon} L^\sigma y_\infty = y_\infty + \eta_\epsilon g^* + \bar{\nu}\left((1-\lambda)^2\epsilon/2\right). \tag{5.12}$$

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

$$L_\pi^k L_{d^\sigma}^\sigma y_\infty = L_\delta^{k-\eta_\epsilon}(L^{\eta_\epsilon} L^\sigma y_\infty) \tag{5.13}$$
$$= y_\infty + kg^* + \bar{\nu}\left((1-\lambda)^2\epsilon/2\right),$$

where the last inequality combines (5.12) with the fact that $\delta \in E(y_\infty)$. Combining (5.11) and (5.13), we deduce that

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

for $\eta_\epsilon < k \le N$ and so (5.9) becomes

$$\|L^k L^\sigma v_N^* - L_\pi^k L_{d^\sigma}^\sigma v_N^*\| \le (1-\lambda)^2\epsilon + \lambda(1-\lambda)\epsilon \tag{5.14}$$
$$= (1-\lambda)\epsilon$$

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

Finally, (5.10) and (5.14) together imply that

$$\|L^k L^\sigma v_N^* - L_\pi^k L_{d^\sigma}^\sigma v_N^*\| \le (1-\lambda)\epsilon \tag{5.15}$$

for $0 \le k \le N$. By Proposition 4.2, it follows that $\{\pi, d^\sigma\}$ is uniformly ϵ-optimal, proving the theorem. □

The next theorem examines the structure of *precisely* optimal policies for large N.

THEOREM 5.3 (STRUCTURE OF OPTIMAL POLICIES). *Assume that Conditions* 1 *and* 3 *hold. Let* $\rho = \Delta(y_\infty)$, (*with* $\Delta(\cdot)$ *as defined in* (5.1)) *and let* $\eta_0 = M_{y_\infty}((1-\lambda)\rho)$. *Then if* $N > \eta_0$, *all uniformly optimal cyclo-stationary policies use decision rules in* $E(y_\infty)$ *on the first* $N - \eta_0$ *fast scale epochs of each slow scale cycle.*

PROOF. Consider an arbitrary uniformly optimal cyclo-stationary policy $\{\pi, d^\sigma\}$. Since $N > \eta_0$, then by Theorem 5.1,

$$v_N^* = \frac{Ng^*}{1-\lambda} + y_\infty + \bar{\nu}(\rho). \tag{5.16}$$

Also, by our definitions of η_0 and y_∞, we have,

$$L^k L^\sigma y_\infty = \hat{L}^\infty L^\sigma y_\infty + kg^* + \bar{\nu}((1-\lambda)\rho) \tag{5.17}$$
$$= y_\infty + kg^* + \bar{\nu}((1-\lambda)\rho)$$

for all $k \ge \eta_0$. Noting that g^* is state-independent, (5.16) and (5.17) may be combined to obtain

$$L^k L^\sigma v_N^* = L^k L^\sigma y_\infty + \frac{\lambda N g^*}{1-\lambda} + \lambda\bar{\nu}(\rho) \tag{5.18}$$
$$= y_\infty + \left(k + \frac{\lambda N}{1-\lambda}\right) g^* + \bar{\nu}(\rho)$$

for all $k \ge \eta_0$. Since $\left(k + \lambda N/(1-\lambda)\right)g^*$ is state-independent, then noting (5.2),

$$\arg\max_{d\in D}\left\{r_d + P_d(L^k L^\sigma v_N^*)\right\} = \arg\max_{d\in D}\left\{r_d + P_d y_\infty + \bar{\nu}(\rho)\right\} \subset E(y_\infty)$$

for all $k \ge \eta_0$. This completes the proof. □

Theorem 5.3 has the following Corollary:

COROLLARY 5.4 (EXISTENCE OF OPTIMAL I.S.P.'s). *Assume that Conditions* 1 *and* 3 *hold and that* $E(y_\infty) = \{\delta\}$ *is a singleton. Let* $\rho = \Delta(y_\infty)$ *and let* $\eta_0 = M_{y_\infty}((1-\lambda)\rho)$. *Then if* $N > \eta_0$, *a uniformly optimal i.s.p. exists with planning horizon* η_0 *and turnpike decision rule* δ.

PROOF. Immediate from Theorem 5.3. □

6. Analysis of the state-dependent gain case. In this section, we allow g^* to be state-dependent. This case is of interest in applications such as multiple project management (see Example 1). In these applications, the projects $\{\Psi^i\}_{i=1}^m$ have optimal gain vectors $\{g_i^*\}_{i=1}^m$. Even when the g_i^* are state-independent over S^i, the g_i^* may be different from one another, in which case the overall gain g^* of Ψ will be state-dependent over S.

In the state-independent gain case, we found, in fairly wide generality, that ϵ-optimal i.s.p.'s exist with gain optimal turnpike decision rules. This lays a foundation for solving the PTMDP (see §9), with effort comparable to that of solving the fast scale average reward MDP. The potential behavior of i.s.p.'s in the state-dependent gain case is quite a bit more varied than the state-independent case. Even in the simplest models, i.s.p.'s with bounded planning horizons may not be ϵ-optimal. Also, i.s.p.'s may require nonaverage optimal turnpike decision rules. These phenomena are illustrated in the following examples.

EXAMPLE 2 (BOUNDED PLANNING HORIZONS IMPLY $N\epsilon$-OPTIMALITY). Consider a PTMDP with

$$S = \{1,2\},\ A_1 = \{1,2\},\ A_2 = \{1\},\ D^\sigma = \{d^\sigma\},\ \lambda = 0.8,$$

fast scale data,

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

and slow scale data

$$r_{d^\sigma}^\sigma = \begin{bmatrix}0\\0\end{bmatrix}, \qquad P_{d^\sigma}^\sigma = \begin{bmatrix}0 & 1\\0.75 & 0.25\end{bmatrix}.$$

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 slow scale cycle and $d_2 \in D$ for the remaining ones. It is immediate to verify that δ is the only gain optimal decision rule and that $g^* = [1\ \ 0]^T$. Hence Condition 1 does not hold. Since all decision rules are aperiodic Condition 3 does hold.

The optimality equation is $v_N^* = L^N L_{d^\sigma}^\sigma v_N^*$. Since state 2 is absorbing with reward zero on the fast scale 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 $[c_N^*\ \ 0.75c_N^*]^T$. Since all rewards are non-negative, $c_N^* > 0$. Optimal policies are obtained via the backward induction sequence,

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

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 $[c_N^\eta \ 0.75c_N^\eta]^T$ and satisfies the fixed point equation,

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

Solving the component equation for state 1, we obtain

$$c_N^\eta = \frac{N-\eta}{0.25+0.15(2^{-\eta})}. \tag{6.1}$$

The optimality problem therefore reduces to finding $\eta^*(N)$, which maximizes c_N^η. Optimizing the right-hand side of (6.1) over η, we obtain that $\eta^*(N)$ is of order $\log_2 N$ and $\lim_{N\to\infty} c_N^*/N = 4$.

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 an i.s.p. will be $N\epsilon$-optimal, not ϵ-optimal, if its planning horizon η_0 is bounded independently of N.

REMARK 7. It can be shown (see Theorem 6.8(b) in Jacobson 1998) that when Condition 1 does not hold,

$$\lim_{N\to\infty} \left\| \frac{v_N^*}{N} \right\|_{\text{sp}} > 0.$$

Hence, in the state-dependent gain case, v_N^* is asymptotically proportional to N. This means that, while $N\epsilon$-optimality is inferior to ϵ-optimality, it is still a reasonable approximation of optimality in a proportional sense.

EXAMPLE 3 (TURNPIKE DECISION RULES THAT ARE NOT GAIN OPTIMAL). Consider a PTMDP with

$$S = \{1, 2, 3\}, \qquad A_1 = \{1, 2\}, \qquad A_2 = \{1\}, \qquad A_3 = \{1, 2\}, \qquad D^\sigma = \{d^\sigma\}, \qquad \lambda = 0.8,$$

fast scale data,

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

where the parameter $r > 0$, and slow scale data

$$r_{d^\sigma}^\sigma = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad P_{d^\sigma}^\sigma = \begin{bmatrix} 0 & 1 & 0 \\ 0.75 & 0.25 & 0 \\ 1-\mu & 0 & \mu \end{bmatrix}.$$

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 turnpike decision rule is not average optimal. In particular, the turnpike 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 2 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. $\text{seg}_2(\delta, \zeta, \eta^*)$ from Example 2. 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 fast scale 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 fast scale 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

$$v_N^*(3) = \max\{0.75c_N^*, 0.8c_N^*(1-\mu)+0.8\mu v_N^*(3) - rN\}. \tag{6.2}$$

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 slow scale cycle is the only possibility in an optimal cyclo-stationary policy and

$$v_N^*(3) = \frac{4(1-\mu)c_N^* - 5rN}{5-4\mu}. \tag{6.3}$$

From (6.2) and (6.3), 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 2, we saw that c_N^* is of the order $4N$. Hence, upon normalizing by N and taking N sufficiently large, the last equality becomes,

$$\lim_{N\to\infty} \frac{v_N^*(3) - 0.75c_N^*}{N} = \frac{1-4\mu-5r}{5-4\mu} \triangleq \kappa(\mu, r). \tag{6.4}$$

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 $N\epsilon$-optimality) is possible for arbitrary $\epsilon \geq 0$ only via an i.s.p. with a non-average optimal turnpike decision rule. Conversely, when $\kappa(\mu, r) < 0$, action 1 is always chosen, implying that a uniform optimal i.s.p. exists with a gain optimal turnpike decision rule.

Example 2 provides a clue about what can be salvaged from the state-independent gain case. It shows that i.s.p.'s whose turnpike decision rule is gain optimal and whose planning horizon is bounded independently of N may still be $N\epsilon$-optimal. Example 3 is a simple counter-example to this, so it is clear that some specific conditions must hold for this to be true. We shall see that Conditions 4 and 5 are sufficient conditions. Note that these conditions depend on the fast scale data only.

For the remainder of this section, we shall assume that Conditions 4 and 5 hold. Since Condition 4 holds, the operator

$$\hat{U}^{\infty}x \stackrel{\triangle}{=} \lim_{k\to\infty} U^k x \tag{6.5}$$

is defined for all $x \in \mathbb{R}^{|S|}$. By analogy with $\hat{L}^{\infty}$, convergence of $\|U^k x - \hat{U}^{\infty}x\|$ to zero is monotonic and geometric with a rate which depends on x. Also, the range of $\hat{U}^{\infty}$ is W and $\hat{U}^{\infty}$ has all the same properties as the operator U. Accordingly, for every vector x and every $\epsilon > 0$, we can define

$$M_U(x,\epsilon) \stackrel{\triangle}{=} \min\left\{k : \|U^k x - \hat{U}^{\infty}x\| \le \epsilon\right\}.$$

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

We now introduce the operator Q,

$$Qx \stackrel{\triangle}{=} g^* + \lambda \hat{U}^{\infty}U^{\sigma}x,$$

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

$$M_{x_{\infty}}(\epsilon) \stackrel{\triangle}{=} M_U(U^{\sigma}\lambda x_{\infty}, \epsilon)$$

which will play the same kind of role as $M_{y_{\infty}}$ in the preceding section.

Our analysis will be aided by the following two propositions.

PROPOSITION 6.1 (DYNAMIC PROGRAMMING ON SCALED VECTORS). *Assume that Conditions* 4 *and* 5 *hold. Let* y_0 *be a terminal reward vector of the form* $y_0 = Nx_0$ *and fix any* $\epsilon > 0$. *Then, the representation,*

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

holds for all $k \ge M_U(x_0,\epsilon)$, *where* c_1, c_2 *are positive constants.*

PROOF. For brevity, let $M = M_U(x_0,\epsilon)$. 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,

$$\begin{aligned} L^M y_0 &= \max_{\pi}\{r_{d_1} + P_{d_1}r_{d_2} + \cdots + P_{\pi}^{M}(Nx_0)\} \\ &= N\max_{\pi}\{P_{\pi}^{M}x_0\} + \bar{\nu}(r_{\max}M) \\ &= NU^M x_0 + \bar{\nu}(r_{\max}M). \end{aligned} \tag{6.6}$$

Recall that $r_{\max}$ is a global bound on the rewards. By the definition of M, we have $U^M x_0 = \hat{U}^{\infty}x_0 + \bar{\nu}(\epsilon)$, so that (6.6) becomes

$$L^M y_0 = N(\hat{U}^{\infty}x_0 + \bar{\nu}(\epsilon)) + \bar{\nu}(r_{\max}M). \tag{6.7}$$

Now from the proof in Theorem 9.4.1 in [8] of the boundedness of $L^n x - ng^*$ we know that for a certain vector h,

$$\begin{aligned} L^n x &= r_{d_1} + P_{d_1}r_{d_2} + \cdots + P_{\pi}^{n}x \\ &\le ng^* + h + P_{\pi}^{n}(x-h) \end{aligned} \tag{6.8}$$

for any given n and 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,

$$L^n x \geq L_\delta^n x = ng^* + h_\delta + P_\delta^n (x - h_\delta) \tag{6.9}$$

where h_δ is the bias vector of δ.

Now, take any $k > M$ and let $x = L^M y_0$ and $n = k - M$. Then (6.8) becomes

$$L^{k-M} L^M y_0 = L^k y_0 \leq (k-M)g^* + h + P_\pi^n L^M y_0 - P_\pi^n h.$$

Upon substituting (6.7), rearranging, and noting that $\hat{U}^\infty x_0 \in W$,

$$\begin{aligned} L^k y_0 &\leq kg^* + N(P_\pi^n \hat{U}^\infty x_0 + \bar{\nu}(\epsilon)) + \bar{\nu}(2\|h\|) + \bar{\nu}(2r_{\max} M) \\ &\leq kg^* + N(\hat{U}^\infty x_0 + \bar{\nu}(\epsilon)) + \bar{\nu}(2\|h\|) + \bar{\nu}(2r_{\max} M). \end{aligned} \tag{6.10}$$

Similarly, (6.9) becomes, once we let $\delta = \gamma$, where γ is as in Condition 5

$$\begin{aligned} L^k y_0 &\geq kg^* + N(P_\gamma^n \hat{U}^\infty x_0 + \bar{\nu}(\epsilon)) + \bar{\nu}\left(2\|h_\gamma\|\right) + \bar{\nu}(2r_{\max} M) \\ &= kg^* + N(\hat{U}^\infty x_0 + \bar{\nu}(\epsilon)) + \bar{\nu}\left(2\|h_\gamma\|\right) + \bar{\nu}(2r_{\max} M). \end{aligned} \tag{6.11}$$

From (6.10) and (6.11), we deduce that

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

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

The following proposition relates $\mathscr{L}_N(Nx_0)$ to Q.

PROPOSITION 6.2 (PTMDP DYNAMIC PROGRAMMING ON SCALED VECTORS). *Assume that Conditions* 4 *and* 5 *hold. Let z_0 be a terminal reward vector of the form $z_0 = Nx_0$ and fix any $\epsilon > 0$. Then, if $N \geq M_U(U^\sigma \lambda x_0, \epsilon)$, the representation*

$$\mathscr{L}_N z_0 = N(Qx_0 + \bar{\nu}(\epsilon)) + \bar{\nu}(c_1 M_U(U^\sigma \lambda x_0, \epsilon)) + \bar{\nu}(c_2) \tag{6.12}$$

holds for positive constants c_1, c_2.

PROOF. Observe that

$$\begin{aligned} L^\sigma z_0 &= \max_{d^\sigma \in D^\sigma} \{r_{d^\sigma}^\sigma + \lambda P_{d^\sigma}^\sigma (Nx_0)\} \\ &= N \max_{d^\sigma \in D^\sigma} \{P_{d^\sigma}^\sigma \lambda x_0\} + \bar{\nu}(r_{\max}) \\ &= NU^\sigma \lambda x_0 + \bar{\nu}(r_{\max}). \end{aligned}$$

By the nonexpansive property of L^N, therefore,

$$\begin{aligned} \mathscr{L}_N z_0 &= L^N L^\sigma z_0 \\ &= L^N (NU^\sigma \lambda x_0) + \bar{\nu}(r_{\max}). \end{aligned} \tag{6.13}$$

Invoking Proposition 6.1 on the first term in (6.13), this becomes

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

from which (6.12) follows. □

The next theorem examines the N-dependence of v_N^*. It shows that 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 5.1, we find that in the state-dependent gain case, the linear term may depend on both the stationary and slow scale data.

THEOREM 6.3 (ASYMPTOTIC BEHAVIOR OF v_N^*, STATE-DEPENDENT GAIN). *Assume that Conditions* 4 *and* 5 *hold. Then*

$$v_N^* = Nx_\infty + O(\log N). \tag{6.14}$$

PROOF. Let $z_0 = Nx_\infty$ and $\epsilon_N = 1/N$. Since value iteration converges geometrically, $M_{x_\infty}(\epsilon_N) = M_{x_\infty}(1/N)$ is $O(\log N)$ and there exists an N_0 such that $N \geq M_{x_\infty}(\epsilon_N)$ for all $N > N_0$. For all $N > N_0$, the hypotheses of Proposition 6.2 are satisfied and,

$$\begin{aligned}\mathscr{L}_N z_0 &= N(Qx_\infty + \bar{\nu}(\epsilon_N)) + \bar{\nu}\left(c_1 M_{x_\infty}(\epsilon_N)\right) + \bar{\nu}(c_2) \\ &= Nx_\infty + O(\log N).\end{aligned}$$

Now, from Theorem 3.2(a) and the Banach Theorem,

$$\begin{aligned}\|v_N^* - z_0\| &\leq \frac{\|\mathscr{L}_N z_0 - z_0\|}{1-\lambda} \\ &= O(\log N)\end{aligned}$$

concluding the proof. □

The following theorem is the main result of this section. It establishes that $N\epsilon$-optimal i.s.p.'s exist whenever Conditions 4 and 5 hold. Moreover, the turnpike decision rule can be any γ satisfying (3.7) and so can be derived directly from Ψ.

THEOREM 6.4 (EXISTENCE OF $N\epsilon$-OPTIMAL I.S.P.'S). *Assume that Conditions* 4 *and* 5 *hold. Let* $\epsilon > 0$, *let* $\eta_\epsilon = M_{x_\infty}((1-\lambda)\epsilon/4)$ *and let* $\gamma \in D^*$ *satisfy* (3.7). *Then for all* N *sufficiently large, a uniform* $N\epsilon$*-optimal i.s.p. exists with planning horizon* η_ϵ *and turnpike decision rule* γ.

PROOF. Let $N > \eta_\epsilon$, let $\epsilon' = (1-\lambda)\epsilon/4$, and let $d^\sigma \in D^\sigma$ be a decision rule satisfying $U^\sigma \lambda x_\infty = U_{d^\sigma}^\sigma \lambda x_\infty$. Let π denote a sequence of N decision rules whose first $N - \eta_\epsilon$ terms are γ and whose remaining terms satisfy $U_\pi^k U_{d^\sigma}^\sigma \lambda x_\infty = U^k U^\sigma \lambda x_\infty$ for $0 \leq k \leq \eta_\epsilon$.

By Theorem 6.3,

$$\begin{aligned}L^\sigma v_N^* &= \max_{d^\sigma \in D^\sigma} \{r_d^\sigma + \lambda P_{d^\sigma}^\sigma (Nx_\infty + O(\log N))\} \\ &= N \max_{d^\sigma \in D^\sigma} \{P_{d^\sigma}^\sigma \lambda x_\infty\} + O(\log N) \\ &= NU^\sigma \lambda x_\infty + O(\log N).\end{aligned} \tag{6.15}$$

Similarly, by the definition of d^σ,

$$L_{d^\sigma}^\sigma v_N^* = NU^\sigma \lambda x_\infty + O(\log N). \tag{6.16}$$

From (6.15),

$$L^k L^\sigma v_N^* = L^k (NU^\sigma \lambda x_\infty) + O(\log N). \tag{6.17}$$

Applying Theorem (6.3) yet again gives, for $0 \le k \le \eta_\epsilon$ and N sufficiently large,

$$\begin{aligned}(6.18)\qquad \|L^k L^\sigma v_N^* - L_\pi^k L_{d^\sigma}^\sigma v_N^*\| &\le N\|U^k U^\sigma \lambda x_\infty - U_\pi^k U_{d^\sigma}^\sigma \lambda x_\infty\| \\ &\quad + 2r_{\max}(\eta_\epsilon + 1) + O(\log N) \\ &= N\|U^k U^\sigma \lambda x_\infty - U_\pi^k U_{d^\sigma}^\sigma \lambda x_\infty\| + O(\log N) \\ &= 0 + (1-\lambda)N\epsilon.\end{aligned}$$

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

When $k > \eta_\epsilon$, Proposition 6.1 applies to the first term on the right hand side of (6.17) which then becomes,

$$\begin{aligned}(6.19)\qquad L^k L^\sigma v_N^* &= kg^* + N(\hat{U}^\infty U^\sigma \lambda x_\infty + \bar{\nu}(\epsilon')) + \bar{\nu}(c_1\eta_\epsilon) + \bar{\nu}(c_2) + O(\log N) \\ &= kg^* + N\hat{U}^\infty U^\sigma \lambda x_\infty + N\bar{\nu}(\epsilon') + O(\log N).\end{aligned}$$

Applying $L_\pi^{\eta_\epsilon}$ to both sides of (6.16), we have

$$\begin{aligned}(6.20)\qquad L_\pi^{\eta_\epsilon} L_{d^\sigma}^\sigma v_N^* &= r_{\max}\bar{\nu}(\eta_\epsilon) + U_\pi^{\eta_\epsilon}(NU^\sigma \lambda x_\infty) + O(\log N) \\ &= \eta_\epsilon g^* + N\hat{U}^\infty U^\sigma \lambda x_\infty + N\bar{\nu}(\epsilon') + O(\log N).\end{aligned}$$

Reinvoking (6.9) with $\delta = \gamma$ and incorporating (6.20),

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

Noting Condition 5, this becomes

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

Subtracting this from (6.19) yields

$$L^k L^\sigma v_N^* - L_\pi^k L_{d^\sigma}^\sigma v_N^* \le N\bar{\nu}((1-\lambda)\epsilon/2) + O(\log N),$$

and, hence,

$$(6.21)\qquad \|L^k L^\sigma v_N^* - L_\pi^k L_{d^\sigma}^\sigma v_N^*\| \le N(1-\lambda)\epsilon$$

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

Taking (6.18) and (6.21) together, it is clear from Proposition 4.2 that the i.s.p. $\{\pi, d^\sigma\}$ is uniformly $N\epsilon$-optimal. This completes the proof. □

REMARK 8. Example 2 is an illustration of Theorem 6.4. Since all decision rules are aperiodic, Condition 4 holds. Furthermore, each state is itself both a Bather class and a Schweitzer-Federgruen class. Consequently Hypotheses 1 and 2 and, hence, also Condition 5, hold. The example therefore satisfies the hypotheses of the theorem and not surprisingly an $N\epsilon$-optimal i.s.p. was observed with $\gamma = \delta$.

Example 3, meanwhile, is a counter-illustration showing what can happen when Condition 5 does not hold. The reader will observe that, in this example, the vector

$w = [1\ \ 1\ \ 0]^T \in W$. However, there is no gain optimal γ satisfying (3.7) for this w. Hence, in contrast to Theorem 6.4, an i.s.p. with a gain optimal turnpike decision rule, may or may not exist, depending on the slow scale data.

REMARK 9. Theorem 6.4 is also applicable to any multiple project management problem (see Example 1) for which the MDPs $\{\Psi^i\}_{i=1}^m$ are all weakly communicating and aperiodic. Here again, aperiodicity implies Condition 4. Since the Ψ^i are weakly communicating, Hypothesis 3, and hence Condition 5, holds as discussed in §3.5. The hypotheses of the theorem are therefore satisfied. The existence of an approximately optimal i.s.p. shows that decisions may be made in a routine, stationary manner for the first part of each slow scale cycle. Toward the end, a planning horizon would be reached during which alternative decision rules would be used, perhaps to close the current project and prepare for a new one.

7. ϵ-Optimality in the state-dependent gain case. In Example 2, we observed that the i.s.p. $\text{seg}_2(\delta, \zeta, \eta^*(N))$ is uniformly optimal and $\eta^*(N)$ is of order $\log_2 N$. This suggests that one might still be able to obtain ϵ-optimal i.s.p.'s in the state-dependent gain case when the planning horizon is an increasing, unbounded function of N. In addition, one might hope that, as in the example,

$$\lim_{N\to\infty} \frac{\eta^*(N)}{N} = 0. \tag{7.1}$$

If so, the planning horizon would be a very small fraction of the slow scale cycle length for sufficiently large N. Initially stationary policies would then have the same benefits as in the constant gain case.

Conditions for the existence of ϵ-optimal i.s.p.'s in the state-dependent gain case were established in Theorem B.5 in Jacobson (1998) for a history-dependent variant of the PTMDP model considered here. The analysis easily generalizes to PTMDPs. In particular, when Hypothesis 3 holds (or alternatively Hypotheses 1 and 2) and the PTMDP is aperiodic, ϵ-optimal i.s.p.'s do exist with planning horizons satisfying (7.1). Multiple project management problems with aperiodic, weakly communicating MDPs Ψ^i therefore satisfy the appropriate conditions (see also Remark 9). At this time, we believe the results can be obtained under still weaker conditions and will consider this topic further in future work.

8. Fully stationary fast scale decision-making. It is natural to wonder what would occur if a single decision rule, δ, were used at all fast scale epochs. This is perhaps the closest analogue to a stationary policy from the standard theory of time-homogeneous MDPs. Since the PTMDP has a nearly time-homogeneous structure when N is large (perturbed only by infrequently occurring slow scale decision epochs), one might wish to know how sub-optimal such decision-making would be in a "nearly time-homogeneous" decision process such as a PTMDP.

Basic bounds on the degree of sub-optimality can of course be obtained from Proposition 4.2. Certain insights can also be made by applying Theorem 5.1 in the state-independent gain case and Theorem 6.3 in the state-dependent gain case. Theorem 5.1 says that the optimal value is asymptotically like $Ng^*/(1-\lambda) + y_\infty$. If one restricts the policy space to those that use a gain optimal δ at all fast scale epochs, then the corresponding optimal value has the asymptotic form $Ng^*/(1-\lambda) + y_\delta$, where y_δ is a modification of y_∞. The asymptotic difference is $y_\infty - y_\delta$. Using Theorem 6.3 in a similar way in the state-dependent case, one finds that the asymptotic degree of sub-optimality is $N(x_\infty - x_\delta) + O(\log N)$.

9. Approaches to policy computation. In Jacobson (1998), generic procedures are suggested for solving the discounted optimality problem for a very similar model to that dealt with here. In this section, we summarize the main points for the model presently considered, and refer the reader to Jacobson (1998) for a more detailed discussion of the

practical issues. In addition, we wish to emphasize that specific insights into the structure of the problem at hand often lead to more efficient solutions than generic procedures.

For the state-independent gain case, observe that, if y_∞ is known up to an additive state-independent vector $c\mathbf{1}$, then $M_{y_\infty}(\cdot)$, $\Delta(y_\infty)$ and $E(y_\infty)$, which appear throughout the analysis of §5, can all be explicitly computed. Accordingly, the i.s.p.'s described in Theorem 5.2 and Corollary 5.4 can be explicitly computed. One situation where it is possible to know y_∞ up to a state-independent vector is when Condition 2 holds. One can then determine a solution $v_0 \in V$ to the average optimality equations using, say, policy iteration. Since $y_\infty \in V$, it follows from Condition 2 that v_0 and y_∞ will differ by some $c\mathbf{1}$. In this case, one can therefore solve the discounted PTMDP optimality problem with the same order of computational effort as when solving the fast scale average reward MDP.

When y_∞ is not known up to an additive state-independent vector, one may try to approximate it by computing the sequence $\{v_n^* - ng^*/(1-\lambda)\}_{n=1}^\infty$. Recall from Theorem 5.1 and Remark 6, that this sequence converges to y_∞ at a geometric rate. To compute the sequence, one can obtain approximations to each v_n^* by iterating the corresponding contraction operator $\mathscr{L}_n$, while g^* can be obtained in a pre-analysis of the fast scale MDP. If $v_{n_0}^* - n_0 g^*/(1-\lambda)$ yields a sufficiently accurate approximation of y_∞ (stopping criteria are suggested in Section 7.5 of Jacobson 1998) and $N \gg n_0$, then this approach is more efficient than the brute force approach of computing $v_N^* - Ng^*/(1-\lambda)$.

For the state-dependent gain case, the i.s.p. of Theorem 6.4 can be explicitly computed if x_∞ is known a priori (by analogy with y_∞). Otherwise, one can again try to iteratively approximate x_∞. An approach to approximating x_∞ suggested in Jacobson (1998) makes use of the operators

$$Q_j x \stackrel{\triangle}{=} g^* + \lambda U^j U^\sigma x, \qquad j = 1, 2, \ldots$$

which are contraction mappings, similar to the Q operator. It can be established (see Theorem 8.21 in Jacobson 1998) that the sequence of fixed points x_j of Q_j converges to x_∞ at a geometric rate. Approximates to each x_j can be obtained by iterating the corresponding Q_j operator.

10. Conclusions. We have introduced a time-inhomogeneous Markov Decision Process model, called a PTMDP, for which we have formulated a discounted optimality problem. Under assumptions on the PTMDP's time-homogeneous, fast scale data alone, we have found that approximately optimal policies exist that exhibit turnpike-like behavior. These so-called initially stationary policies have turnpike decision rules that are gain optimal. Furthermore, the size of their planning horizon is independent of N. For large N, this constitutes a substantial simplification in structure as compared to more general policies. In the course of our analysis, we also characterized the form of the optimal value vector (see Theorems 5.1 and 6.3).

In the case where the fast scale MDP had state-independent gain, ϵ-optimal i.s.p.'s exist under fairly weak assumptions. Beyond the assumption of state-independent gain itself, we have applied Condition 3. However, this mainly simplifies the presentation of our analysis. When Condition 3 does not hold, generalizations of our analysis that account for periodicity effects are possible, although trite. When Condition 2, a condition of common interest in the literature, is applied, an ϵ-optimal i.s.p. can be explicitly computed with minimal effort. More complicated procedures are necessary when Condition 2 does not hold.

In the case where the fast scale MDP had state-dependent gain, we imposed Conditions 4 and 5. The latter condition is not a general one, however, it does hold in certain applications of interest, such as multiple project management problems. Under these conditions, i.s.p.'s with the same structural properties as in the state-independent gain case were shown to exist. However, these i.s.p.'s were $N\epsilon$-optimal rather than ϵ-optimal. This scaled ϵ-optimality is reasonable, however, since the optimal value also scales with N.

It has been shown in related work that ϵ-optimal i.s.p.'s are possible in the state-dependent gain case by letting the planning horizon grow as a function of N. Sufficient conditions for this are satisfied by certain multiple project management problems whose projects can be modeled by aperiodic, weakly communicating MDPs. In future work, we will attempt to find still weaker sufficient conditions.

Acknowledgments. The authors thank the referees for many helpful suggestions for improving the presentation of this paper. This work was done while M. Jacobson was with the Department of Electrical Engineering, Technion, Israel Institute of Technology.

References

Bather, J. 1973. Optimal decision procedures for finite Markov chains III. *Adv. Appl. Probab.* **5** 541–553.

Bäuerle, N. 2001. Discounted stochastic fluid programs. *Math. Oper. Res.* **26** 401–420.

Davis, M. H. A. 1993. *Markov Models and Optimization*. Chapman and Hall, London, U.K.

Delebecque, F., J. P. Quadrat. 1981. Optimal control of Markov chains admitting strong and weak interactions. *Automatica* **17** 281–296.

Hinderer, K., G. Hubner. 1977. On approximate and exact solution for finite stage dynamic programs. H. Tijms, J. Wessels, eds. *Markov Decision Theory.* Mathematical Centre, Amsterdam, The Netherlands.

Jacobson, M. 1998. Asymptotic properties of two timescale Markov decision processes. M.Sc. thesis, Technion, Israel Institute of Technology, Haifa, Israel. (Available at http://www.ee.technion.ac.il/~adam/PAPERS/MWJ-MSc.ps).

Phillips, R. G., P. Kokotovic. 1981. A singularly perturbation approach to modeling and control of Markov chains. *IEEE Trans. Automatic Control* **AC-26** 1087–1094.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York.

Schweitzer, P. J. 1984. A value iteration scheme for undiscounted multichain markov renewal programs. *Zeitschrift Oper. Res.* **28** 143–152.

———, A. Federgruen. 1977. The asymptotic behavior of undiscounted value iteration in Markov decision problems. *Math. Oper. Res.* **2** 360–382.

———, ———. 1978. The functional equations of undiscounted Markov renewal programming. *Math. Oper. Res.* **3** 308–321.

———, ———. 1979. Geometric convergence of value iteration in multichain Markov decision processes. *Adv. Appl. Probab.* **11** 188–217.

Sethi, S. P., Q. Zhang. 1994. *Hierarchical Decision Making in Stochastic Manufacturing Systems*. Birkhäuser, Boston, MA.

Shapiro, J. 1968. Turnpike planning horizons for a Markovian decision model. *Management Sci.* **14** 292–300.

M. Jacobson: Department of Electrical Engineering and Computer Science. The University of Michigan, Ann Arbor, MI 48109; e-mail: mwjacobs@eecs.umich.edu

N. Shimkin: Department of Electrical Engineering, Technion, Israel Institute of Technology; e-mail: shimkin@ee.technion.ac.il

A. Shwartz: Department of Electrical Engineering, Technion, Israel Institute of Technology; e-mail: adam@ee.technion.ac.il