Proceedings of the 34th
Conference on Decision & Control
New Orleans, LA - December 1995

TP05 3:00

Weighted Markov Decision Processes with Perturbation*

Ke Liu[†] and Jerzy A. Filar[‡]
School of Mathematics
University of South Australia, The Levels
Adelaide, South Australia 5095
kliu@rincewind.levels.unisa.edu.au/j.filar@unisa.edu.au

Abstract

In this paper we consider the weighted reward MDP's with perturbation. We give the proof of existence of a δ-optimal simple ultimately deterministic policy under the assumption of "scalar value". We also prove that there exists a δ-i-optimal simple ultimately deterministic policy in the perturbed weighted MDP, for all $\epsilon \in [0, \epsilon^*)$ even without the assumption of "scalar value".

1. Introduction

A discrete Markov Decision Process (MDP, for short) Γ can be informally described as follows: At time points $t = 0, 1, 2, \ldots$, a process Γ is observed and decision maker finds himself in some *state* i with a choice of available *actions*. Choosing a particular action, a, results in two things: (i) an immediate reward $r(i, a)$ is accrued which depends on the current state and the action chosen, and, (ii) the process Γ moves to state j, with *transition probability* $p(j \mid i, a)$ depending on the current state, destination state and the chosen action. The successive immediate rewards obtained during the *infinite time horizon* are aggregated according to some *overall reward criterion* (e.g. future rewards might be discounted). The goal is to choose a *policy* (a course of action) that maximizes the overall reward criterion.

Let S be the finite state space and $N =\mid S \mid$, $A(i)$ the finite set of actions available at state i, and $A = \cup\{A(i) \mid i \in S\}$. Then Γ is synonymous with the four-tuple: $\Gamma = \langle S, A, p, r\rangle$, where p is the family of transition probabilities $p = \{p(j \mid i, a) : (i, a, j,) \in S \times A(i) \times S\}$, and r is the collection of rewards $r = \{r(i, a) : (i, a) \in S \times A(i)\}$. A *decision rule*, π_t at time t, is a function which assigns a probability to the event that any particular action a is taken at time t. In general π_t may depend on all realized states up to and including time t, and on all realized actions up to time $t - 1$.

Let (X_t, Y_t) be the random variables representing respectively the state and action at time t. Let $h_t = (i_0, a_0, i_1, a_1, \ldots, a_{t-1}, i_t)$ be the *history* up to time t, where $a_0 \in A(i_0), \ldots, a_{t-1} \in A(i_{t-1})$, and $i_k \in S, k = 0, 1, \ldots, t$, then $\pi_t(\cdot \mid h_t)$ is a probability distribution on $A(i_t)$, that is $\pi_t(a_t \mid h_t)$ is the probability of selecting the action a_t at time t, given the history h_t. A *policy* π is a sequence of decision rules $\pi = (\pi_0, \pi_1, \ldots, \pi_t, \ldots)$. A *Markov policy* is one in which π_t depends only on the "current" state at time t, i_t. A *stationary policy* π which can be denoted by $\pi = (\pi_0^\infty)$ is a Markov policy with identical decision rules. A *deterministic policy* π is a stationary policy whose single decision rule is nonrandomized.

Let $C, C(M), C(S)$ and $C(D)$ denote the sets of all policies, all Markov policies, all stationary policies and all deterministic policies respectively. We note that a stationary policy $\pi \in C(S)$ can be defined by the vector: $\pi = (\pi(a \mid i); (i, a) \in S \times A(i))$, where $\pi(a \mid i)$ is the probability that the decision maker chooses action $a \in A(i)$ in state i whenever that state is visited. Of course, we have $\pi(a \mid i) \geq 0, a \in A(i)$ and $\sum_{a \in A(i)} \pi(a \mid i) = 1$ for all $i \in S$.

Hence a deterministic policy $\pi \in C(D)$ satisfies $\pi(a \mid i) \in \{0, 1\}$ for all $(i, a) \in S \times A(i)$, and in this case π can be regarded as the map defined from S to A by: $\pi(i) = a \iff \pi(a \mid i) = 1$. For notational convenience, $\pi \in C(D)$ is denoted by f^∞ or simply f, where f is the mapping defined by $f(i) = \pi(i)$.

A stationary policy $\pi = (\pi_0^\infty)$ induces a finite Markov Chain with state space S and the transition matrix $P(\pi_0)$, where for all $t > 1$

$$\begin{aligned} P(\pi_0)_{ij} &= P_\pi(X_t = j \mid X_{t-1} = i) \\ &= \sum_{a \in A(i)} \pi_0(a \mid i) p(j \mid i, a). \end{aligned}$$

Definition 1.1 *A policy π is "ultimately stationary" if there exists a stationary policy $(\pi_0^{'\infty})$ called the "tail" of π, and a random stopping time τ called the "switching time" of π, such that: 1. $\pi_t = \pi_0{}'$ for any $t \geq \tau$, 2. $P_\pi(\tau < \infty) = 1$, where P_π is the distribution induced by the policy π.*

Let $C(US)$ be the set of all ultimately stationary policies. A simpler subclass of $C(US)$ is defined by:

[*] This research was supported partially by National Natural Science Foundation of China.

[†] On temporary leave from Institute of Applied Mathematics, Academia Sinica

[‡] Professor of Mathematics and Statistics, University of South Australia

0-7803-2685-7/95 $4.00 © 1995 IEEE

Definition 1.2 *A policy π is a "simple ultimately deterministic policy" if: 1. π is ultimately stationary and Markov, 2. the tail of π is deterministic, 3. the switching time τ is fixed (i.e., nonrandom). This class of policies will be denoted by C(SUD)*

Let random variable R_t denote the immediate reward at time t. Then for any policy π and initial state i, the expectation of R_t is given by

$$\begin{aligned} &E_\pi(R_t, i) \\ &= \sum_{j\in S}\sum_{a\in A(j)} P_\pi(X_t = j, Y_t = a \mid X_0 = i) r(j,a). \end{aligned}$$

The manner in which the resulting stream of expected rewards $\{E_\pi(R_t,i) \mid t = 0, 1, \dots\}$ is aggregated defines the MDP's discussed in the sequel.

Discounted MDP's: For a policy $\pi \in C$ and an initial state $i \in S$, this overall reward criterion is defined by

$$V_\beta(\pi; i) = \sum_{t=0}^{\infty} \beta^t E_\pi(R_t, i); \qquad i \in S, \tag{1}$$

where $\beta \in [0,1)$ is a fixed discount factor.

We will denote the process by $\Gamma(\beta)$. This is probably the most popular overall payoff criterion for MDP's. Not only is it attractive in an economic setting, but it also avoids the difficulty with the total reward criterion (clearly $V_\beta(\pi;i)$ is finite for any π). We will denote by $V_\beta(\pi)$ the vector whose i-th entry is $V_\beta(\pi;i)$. Note that discounting assigns more weight to early rewards than to later ones.

Average Reward MDP's: For a policy $\pi \in C$ and an initial state $i \in S$, this overall reward criterion is defined by

$$\bar V(\pi; i) = \lim_{T\to\infty} \inf \frac{1}{T+1}\sum_{t=0}^{T} E_\pi(R_t, i); \qquad i \in S. \tag{2}$$

We will denote the process by Γ(A). This "long-run average" criterion is also quite popular. The latter is useful in situations where discounting is not appropriate (e.g., some engineering applications). We note that the average reward criterion "ignores" the rewards earned during any finite time period. We will denote by $\bar V(\pi)$ the vector whose i-th entry is $\bar V(\pi;i)$. Since S and $A(i)$ for $i \in S$ are finite sets and

$$\min_{a\in A(i), i\in S} r(i,a) \le E_\pi(R_t, j) \le \max_{a\in A(i), i\in S} r(i,a), \tag{3}$$

the reward functions of (1) and (2) are bounded.

Definition 1.3 *For $\Gamma(\beta)$, a policy π^* is called "discount optimal", if for all $i \in S$,*

$$V_\beta(\pi^*; i) = \max_{\pi\in C} V_\beta(\pi; i) = V_\beta^*(i), \tag{4}$$

where $V_\beta^*(i)$ is called the i-th entry of "value vector" of $\Gamma(\beta)$, and V_β^* is the value vector of $\Gamma(\beta)$.

Definition 1.4 *For $\Gamma(A)$, a policy π^* is called "average optimal", if for all $i \in S$,*

$$\bar V(\pi^*; i) = \max_{\pi\in C} \bar V(\pi; i) = \bar V^*(i), \tag{5}$$

where $\bar V^*(i)$ is called the i-th entry of value vector of $\Gamma(A)$, and $\bar V^*$ is the value vector of $\Gamma(A)$.

Weighted Reward Criterion: For a policy $\pi \in C$ and an initial state $i \in S$, the overall reward criterion is defined by

$$\omega(\pi; i) = \lambda(1-\beta)V_\beta(\pi; i) + (1-\lambda)\bar V(\pi; i), \tag{6}$$

where $\lambda \in [0, 1]$ is a fixed weight parameter, and β is the discount factor in the process $\Gamma(\beta)$.

We will denote the process by $\Gamma(\beta, \lambda)$. This, and related criteria, have been discussed extensively in recent literature (Krass [14], Ghosh and Marcus [11], Fernandez-Gaucherand, Ghosh and Marcus [8], Filar and Vrieze [10], Feinberg and Shwartz [7] and so on). We will denote by $\omega(\pi)$ the vector whose i-th entry is $\omega(\pi;i)$.

Definition 1.5 *For $\Gamma(\beta,\lambda)$, a policy π^* is called "weighted optimal", if for all $i \in S$,*

$$\omega(\pi^*; i) = \max_{\pi\in C} \omega(\pi; i). \tag{7}$$

A policy π^* is called δ-optimal, if for all $i \in S$,

$$\omega(\pi^*; i) \ge \sup_{\pi\in C} \omega(\pi; i) - \delta, \tag{8}$$

where δ is a nonnegative constant.

2. Basic Results for MDP's

Since the policies in C can be extremely complicated, a central idea in the theory of MDP's has been to localize the search for an optimal policy to "simpler" subclasses of policies. The following result due to Derman and Strauch [6] allows us to consider only policies in $C(M)$.

Theorem 2.1 *For any policy $\pi \in C$ there exists $\pi' \in C(M)$ such that for $t = 0, 1, 2, \dots$,*

$$P_{\pi'}(X_t = j, Y_t = a \mid X_0 = i) = P_\pi(X_t = j, Y_t = a \mid X_0 = i)$$
$$\forall a \in A(j), \quad \forall i, j \in S. \tag{9}$$

This implies that for any optimizing criterion based on the probabilities in (9), we may restrict ourselves to Markov policies. We will use some basic results for $\Gamma(\beta)$ and $\Gamma(A)$. Most of the results are "classical" and can be found in any good reference on MDP's (e.g. Puterman [16], Ross [17], Kallenberg [12], Derman [5], Blackwell [3], Mine and Osaki [15], etc.).

Theorem 2.2 *For any stationary policy $\pi = (\pi_0{}^{\infty}) \in C(S)$, we have:*

$$\begin{aligned} \mathbf{V}_\beta(\pi) &= \mathbf{r}(\pi_0) + \beta P(\pi_0)\mathbf{V}_\beta(\pi), & (10) \\ \mathbf{V}_\beta(\pi) &= [I_N - \beta P(\pi_0)]^{-1}\mathbf{r}(\pi_0), & (11) \end{aligned}$$

where I_N is the identity matrix of dimension N.

For any deterministic Markov policy $\pi = (f_0, f_1, \ldots)$ (10) has the form $V_\beta(\pi) = r(f_0) + \beta P(f_0)V(\pi(1))$; where $\pi(1) = (f_1, f_2, \ldots)$ is a deterministic Markov policy too.

Theorem 2.3
(i) The discounted value vector V_β^* is the unique solution to the following so-called discounted optimality equations (one for each $i \in S$)

$$V(i) = \max_{a \in A(i)} \left\{ r(i,a) + \beta \sum_{j \in S} p(j \mid i,a)V(j) \right\}. \quad (12)$$

(ii) Let $a(i) \in A(i)$ be such that for all $i \in S$

$$a(i) = arg \max_{a \in A(i)} \left\{ r(i,a) + \beta \sum_{j \in S} p(j \mid i,a)V(j) \right\}. \quad (13)$$

Then the deterministic policy π^* defined by: $\pi^*(i) = a(i)$, $i \in S$ is discount optimal.

Definition 2.1 *MDP Γ is called "unichain" if for any deterministic policy $f \in C(D)$, the Markov chain induced by $P(f)$ has one ergodic set plus a (perhaps empty) set of transient states.*

A set of states B "communicates" if for any $i, j \in B \subseteq S$, there exists a policy $\pi \in C(S)$ such that $P_\pi(X_t = i \mid X_0 = j) > 0$ for some t. MDP Γ is called "communicating" if the state space S communicates. An MDP with general structure is called "multichain".

For any deterministic stationary policy $f^\infty \in C(D)$, let

$$P^*(f) = \lim_{n \to \infty} \frac{1}{n} \sum_{m=0}^{n-1} P^m(f). \quad (14)$$

Theorem 2.4 *There exists $\beta^* \in (0,1)$ and $f^* \in C(D)$ such that:1. f^* is discount-optimal policy for all $\beta \in (\beta^*, 1)$; 2. f^* is average optimal policy.*

Proof: See Blackwell [3]. □

Theorem 2.5 *If $\Gamma(A)$ is communicating or unichain then: (i) $\bar{V}^*(i) = \bar{V}^*(j)$ for all $i, j \in S$, that is, the average reward of any optimal policy is independent of the starting state; (ii) There exists an unichain policy $f \in C(D)$ such that $\bar{V}(f; i) = \bar{V}^*(i)$ for all $i \in S$.*

Proof: (i) See Filar and Schultz [9].
(ii) By theorem 2.4 there exists an optimal policy $f \in C(D)$. The state space of Markov process induced by f is $S_1(f) \cup S_2(f) \cup \ldots \cup S_{m(f)}(f) \cup T(f)$, where $S_1(f), S_2(f), \ldots, S_{m(f)}(f)$ are ergodic classes and $T(f)$ is the transient class. Now we have two cases:
Case 1 If $m(f) = 1$, the theorem is proved,
Case 2 If $m(f) > 1$, let $R(f)$ denote the number of elements in the set $S_2(f) \cup \ldots \cup S_{m(f)}(f)$ and consider the action set:

$$\begin{aligned} A^*(f) = \cup_{i \in S \setminus S_1(f) \cup T(f)} \{ & a \in A(i) \\ & | p(S_1(f) \cup T(f) | i, a) > 0 \}. \end{aligned} \quad (15)$$

If $A^*(f) = \emptyset$, then this is a contradiction with the communicating assumption. So define:

$$f_1 = \begin{cases} f(i); & \text{if } i \in S_1(f) \cup T(f) \\ f(i); & \text{if } i \in S \setminus \{S_1(f) \cup T(f)\} \\ & \text{and } A(i) \cap A^*(f) = \emptyset \\ a; & \text{if } i \in S \setminus \{S_1(f) \cup T(f)\} \\ & \text{and } \forall a \in A(i) \cap A^*(f) \neq \emptyset \end{cases} \quad (16)$$

The state space of Markov process induced by f_1 is $S_1(f_1) \cup S_2(f_1) \cup \ldots \cup S_{m(f_1)}(f_1) \cup T(f_1)$, where $S_1(f_1) = S_1(f)$, $T(f_1) \supset T(f)$, $R(f_1) < R(f)$ and for all $i \in S_1(f_1) \cup T(f_1)$ we have $\bar{V}(f_1; i) = \bar{V}^*(i)$. We can define $A^*(f_1)$ as in (15) and continue our steps. Finally, because of the finiteness of state space and action set, we must stop in case 1 after a finite number of steps, then the theorem is proved. □

Models with the scalar value property often occur in practice and have received a lot of attention. For the weighted MDP, the following results can be found in Krass, Filar and Sinha [13] and elsewhere. Note that theorem 2.1 implies: $\sup_{C(M)} \omega(\pi; i) = \sup_C \omega(\pi; i)$, for all $i \in S$, so we need only consider Markov policies. From Krass [14] we have the following results:

Theorem 2.6 *Consider the weighted process $\Gamma(\beta, \lambda)$, with $\lambda > 0$. Suppose that the value $\bar{V}^*(i) \equiv \bar{V}^*$ for all $i \in S$, then: (i) $\sup_C \omega(\pi; i) = \lambda(1-\beta)V_\beta^*(i) + (1-\lambda)\bar{V}^*$ for all $i \in S$. (ii) Given any $\delta > 0$, there exists $\pi(\delta) \in C(SUD)$ which is δ-optimal in $\Gamma(\beta, \lambda)$. (iii) Let $f_0^* \in C(D)$ be optimal in $\Gamma(A)$ and $g_0^* \in C(D)$ be optimal in $\Gamma(\beta)$ and $\delta > 0$ be given. There exists (non-random) $\tau(\delta)$ such that $\pi(\delta) = (f_0, f_1, \ldots)$, with $f_t = f_0^*$ for $t \leq \tau(\delta)$ and $f_t = g_0^*$ for $t > \tau(\delta)$, is δ-optimal in $\Gamma(\beta, \lambda)$.*

Definition 2.2 *For any $\delta > 0$ a policy π is an δ-i-optimal policy if $\omega_i(\pi) \geq \sup_{C(M)} \omega_i(\pi) - \delta$ for all $i \in S$.*

Theorem 2.7 *Consider the weighted process $\Gamma(\beta, \lambda)$ with $\lambda > 0$. Let $g_0^* \in C(D)$ be optimal in $\Gamma(A)$. Given a fixed initial state i and arbitrary $\delta > 0$, there exists a positive non-random $\tau(\delta)$ and a policy $\pi(\delta) = (f_0, f_1, \ldots) \in C(SUD)$ with $f_t = g_0^*$ for all $t > \tau(\delta)$ such that $\pi(\delta)$ is δ-i-optimal in $\Gamma(\beta, \lambda)$.*

Remark 2.1 From Krass [14] we know that the policy $\pi(\delta)$ which is in theorem 2.6 and theorem 2.7 is a Markov policy. Actually from theorem 2.8 we know that the policy $\pi(\delta)$ is a deterministic Markov policy.

The idea of theorem 2.7 is the following: in the weighted MDP, try to maximize the discounted payoff in the early stages and to maximize the average payoff in the later stages. Now we can construct an algorithm for the δ-optimal simple ultimately deterministic policy (see definition 1.2) in the weighted process $\Gamma(\beta, \lambda)$.

Definition 2.3 *Let $\Gamma(\beta, \lambda_1, \lambda_2)$ process is the MDP with overall reward criterion defined for $i \in S$ by*

$$\omega[\lambda_1, \lambda_2](\pi; i) = \lambda_1(1-\beta)V_\beta(\pi; i) + \lambda_2\bar{V}(\pi; i), \quad (17)$$

where $\lambda_1, \lambda_2 \in [0,1]$, β is the discount factor in the process $\Gamma(\beta)$ and $\pi \in C$. Note that the weighted reward process is denoted by $\Gamma(\beta, \lambda, (1-\lambda))$.

Theorem 2.8 *For fixed $\delta > 0$. Choose integer $N = \max_{i \in S} N_i$, where N_i is the smallest positive integer such that $\beta^{N+1}\lambda_1(1-\beta)V_\beta^*(i) \leq \delta$. Let f^* be an average-optimal deterministic policy. Set $f = f^*$.*

```
For $k = N$ down to 0,
    For each $i \in S$, select $a_i$ that achieves
    $\max_{a \in A(i)}\{r(i,a)(1-\beta)\lambda_1\beta^k +$
    $\sum_j p(j|i,a)\omega[\beta^{k+1}\lambda_1, \lambda_2](f; j)\}$.
    Let $f^k$ be the non-randomized decision rule
    taking action $a_i$ in state $i$.
    Set $f = (f^k, f)$ again, decrement $k$, and
    repeat.
```

Now $f = (f^0, f^1, \ldots, f^N, f^*, f^*, \ldots)$ constructed above is an δ-optimal policy in $\Gamma(\beta, \lambda_1, \lambda_2)$.

Theorem 2.9 *There exist $\beta^* \in (0,1)$ and $f^* \in C(D)$ such that: for all $\beta \in [\beta^*, 1)$ and $\lambda \in [0,1]$, f^* is weighted-optimal policy.*

Proof. By theorem 2.4, there exist $\beta^* \in (0,1)$ and $f^* \in C(D)$ such that: (1) f^* is discount-optimal for all $\beta \in [\beta^*, 1)$; and (2) f^* is an average-optimal policy. So for all $\lambda \in [0,1]$ and $\beta \in [\beta^*, 1)$, we have

$$\lambda(1-\beta)V_\beta(f^*; i) = \sup_C \lambda(1-\beta)V_\beta(\pi; i); i \in S \quad (18)$$

$$(1-\lambda)\bar{V}(f^*; i) = \sup_C (1-\lambda)\bar{V}(\pi; i); i \in S. \quad (19)$$

Substitute (18) and (19) into (17), for all $i \in S$ we have

$$\begin{aligned} &\omega(f^*; i) \\ &= \lambda(1-\beta)V_\beta(f^*; i) + (1-\lambda)\bar{V}(f^*; i) \\ &= \max_{\pi \in C} \lambda(1-\beta)V_\beta(\pi; i) + \max_{\pi \in C}(1-\lambda)\bar{V}(\pi; i) \\ &\geq \max_{\pi \in C} \omega(\pi; i) \geq \omega(f^*; i). \quad (20) \end{aligned}$$

Therefore, $\omega(f^*; i) = \max_{\pi \in C} \omega(\pi; i)$. □

3. MDP Perturbation Theory

Consider the MDP Γ and let $P(f)$ be the one step transition probability matrix under the control of stationary policy $f \in C(D)$. When we consider the situation where the transition probabilities of Γ are perturbed slightly, there are now many results. For a recent survey of some of these we refer the reader to Abbad and Filar [2].

Singular Perturbation

Definition 3.1 *A set D is called the disturbance law if: $D = \{d(j \mid i, a) \mid \; i, j \in S, \; a \in A(i)\}$, and the elements of D satisfy: (i) $\sum_{j \in S} d(j \mid i, a) = 0$ for all $i \in S$, $a \in A(i)$, and, (ii) there exists $\epsilon_0 > 0$ such that for all $\epsilon \in [0, \epsilon_0]$, $i \in S$, and $a \in A(i)$,*

$$p(j \mid i, a) + \epsilon d(j \mid i, a) \geq 0. \quad (21)$$

Note that D is more general than the perturbation permitted by Delebecque [4] where it is assumed that $d(j \mid i, a) \geq 0$ whenever $j \neq i$.

Now we have a family of perturbed MDP Γ_ϵ for all $\epsilon \in [0, \epsilon_0]$ that differ from the original MDP Γ only in the transition law, namely, in Γ_ϵ we have that for all $i, j \in S$ and $a \in A(i)$, $p(j \mid i, a)(\epsilon) = p(j \mid i, a) + \epsilon d(j \mid i, a)$. Let $P(\epsilon)$ denote the one step transition probability matrix $P + \epsilon D$, and for the model Γ_ϵ let $P(f, \epsilon) = P(f) + \epsilon D(f)$; $P(\pi_0, \epsilon) = P(\pi_0) + \epsilon D(\pi_0)$, where $D(f)$ and $D(\pi_0)$ are disturbance matrix corresponding to the policies $f \in C(D)$ and $\pi_0 \in C(S)$ respectively.

Definition 3.2 *For $\pi \in C$ and $i \in S$ discounted criterion $V_\beta(\pi; i)(\epsilon)$ and average criterion $\bar{V}(\pi; i)(\epsilon)$ are defined in the same way as $V_\beta(\pi; i)$ and $\bar{V}(\pi; i)$ were defined in $\Gamma(\beta)$ and $\Gamma(A)$ with one step transition probability $P(\pi, \epsilon)$ respectively. Let $V_\beta(\pi, \epsilon)$ and $\bar{V}(\pi, \epsilon)$ denote the vector forms of $V_\beta(\pi; i)(\epsilon)$ and $\bar{V}(\pi, \epsilon)$ respectively.*

Let $\Gamma_\epsilon(\beta)$ and $\Gamma_\epsilon(A)$ be defined in the same way as $\Gamma(\beta)$ and $\Gamma(A)$ respectively and $V_\beta^*(\epsilon) = \max_{\pi \in C} V_\beta(\pi, \epsilon)$; $\bar{V}^*(\epsilon) = \max_{\pi \in C} \bar{V}(\pi, \epsilon)$. For every policy $\pi \in C$, we define $V_\beta(\pi)(0) = \lim_{\epsilon \to 0} V_\beta(\pi)(\epsilon)$; $\bar{V}(\pi)(0) = \liminf_{\epsilon \to 0} \bar{V}(\pi)(\epsilon)$. The following three results follow from Abbad and Filar [2].

Theorem 3.1 *There exists a deterministic policy $f \in C(D)$ and a positive number ϵ_1 such that for any $\epsilon \in [0, \epsilon_1)$, f is a maximizer in $\Gamma_\epsilon(\beta)$.*

Lemma 3.1 *For any stationary policy $\pi \in C(S)$, there exists a "limit stationary matrix": $P_0^*(\pi) = \lim_{\epsilon \to 0} P_\epsilon^*(\pi)$, where $P_\epsilon^*(\pi)$ is defined by (14) with $P(\pi, \epsilon)$.*

Theorem 3.2 *There exists a deterministic policy $f \in C(D)$ and a positive number ϵ_2, such that for any $\epsilon \in (0, \epsilon_2)$, f is a maximizer in $\Gamma_\epsilon(A)$. Moreover, f is a maximizer in $\Gamma_0(A)$.*

Remark 3.1 Theorems 3.1 and 3.2 can be extended to the case where the perturbation is of the form $P + D(\epsilon)$ where all elements of the disturbance law $D(\epsilon)$ are rational functions of ϵ (see Abbad [1]).

General Perturbation

For general perturbation, we have the transition probabilities as follows for $i, j \in S$; $a \in A(i)$

$$p(j \mid i, a)(d) = p(j|i, a) + d(j|i, a), \qquad (22)$$

where the modulus $\|D\| = d = \max\{|d(j|i,a)| \; |a \in A(i);\ i, j \in S\}$ of disturbance law D is small enough $(\|D\| = d < \epsilon_0)$ so that (22) is a transition probability, that is , for any $a \in A(i)$ and $i, j \in S$, $p(j|i,a)(d) \geq 0$ and $\sum_{j \in S} p(j|i,a)(d) = 1$.

It is known that in general $P^*(\pi, d)$ may not have a limit when d tends to 0. For the general disturbance law, Abbad [1] derives the following results.

Theorem 3.3 *Let $\pi \in C(S)$ be any maximizer in the $\Gamma(\beta)$. Then for all $\delta > 0$, there exists $\epsilon(\beta) > 0$ such that for all disturbance law D satisfying $d < \epsilon(\beta)$, $\|V_\beta(\pi, d) - V_\beta^*(d)\| < \delta$.*

Lemma 3.2 *For the $\Gamma(A)$ we have (i) Let $\pi \in C(S)$ be unichain, then we have $\lim_{d \to 0} P^*(\pi, d) = P^*(\pi)$; (ii) Let $\pi = (f_0, f_1, \ldots, f_\tau, f, \ldots) \in C(SUD)$ where $f_\tau = f$, $\tau < \infty$ and f is unichain, then we have $\lim_{d \to 0} P^*(\pi, d) = P^*(\pi) = P^*(f)$.*

Theorem 3.4 *(i) Assume that $\Gamma(A)$ is unichain. Let $\pi \in C(S)$ be any maximizer in the $\Gamma(A)$. Then for all $\delta > 0$, there exists $\epsilon(\delta) > 0$ such that for all disturbance law D satisfying $d < \epsilon(\delta)$, $\|\bar{V}(\pi, d) - \bar{V}^*(d)\| \leq \delta$.*

(ii) Assume that $\Gamma(A)$ is communicating. Let $\pi \in C(S)$ be any maximizer unichain policy in the $\Gamma(A)$. Then for all $\delta > 0$, there exists $\epsilon(\delta) > 0$ such that for all disturbance law D satisfying $d < \epsilon(\delta)$, $\|\bar{V}(\pi, d) - \bar{V}^*(d)\| \leq \delta$.

4. Perturbed Weighted Criterion

Now we will consider the cases with the disturbance law $D(\epsilon)$ which is mentioned in remark 3.1.

Consider the weighted reward MDP(WMDP for short) $\Gamma(\beta, \lambda)$ with perturbation, denoted by $\Gamma_\epsilon(\beta, \lambda)$, which is defined by: $\Gamma_\epsilon(\beta, \lambda) = \langle S, A, p_\epsilon, r, \omega \rangle$, where the perturbed transition law, p_ϵ, is defined by $p_\epsilon = \{p(j|i,a)(\epsilon) : (i, a, j) \in S \times A(i) \times S\}$, and $\omega(\epsilon)$ is given in (6) with p_ϵ and $p(j|i,a)(\epsilon)$ given in (21). Thus for any policy $\pi \in C$, we have for all $i \in S$:

$$\begin{aligned}&\omega(\pi; i)(\epsilon)\\ &\quad = \lambda(1-\beta)V_\beta(\pi; i)(\epsilon) + (1-\lambda)\bar{V}(\pi; i)(\epsilon); \quad (23)\end{aligned}$$

and $\omega^*(i)(\epsilon) = \sup_{\pi \in C} \omega(\pi; i)(\epsilon)$, $i \in S$. For every policy $\pi \in C$, we have $\omega(\pi; i)(0) = \liminf_{\epsilon \to 0} \omega(\pi; i)(\epsilon)$, $i \in S$.

Definition 4.1 *The optimization problem:*

$$\omega^*(i)(0) = \sup_{\pi \in C} \omega(\pi; i)(0), \;\; i \in S \qquad (24)$$

is called the "Limit Weighted Reward MDP" (LWMDP for short).

5. The Properties of LWMDP

In this section we shall attempt to develop the theory for the limit weighted reward MDP's $\Gamma_0(\beta, \lambda)$ introduced in section 4 and for the asymptotics of the $\Gamma_\epsilon(\beta, \lambda)$ as $\epsilon \to 0$. Our approach is along the lines of the "classical" development for the discounted and average MDP's and their perturbed models. For any $\epsilon \in [0, \epsilon_0)$ it is easy to find the upper bound of our model:

$$\begin{aligned}&\sup_{f \in C(D)} \omega(f; i)(\epsilon) \leq \sup_{\pi \in C(S)} \omega(\pi; i)(\epsilon) \leq \sup_{\pi \in C} \omega(\pi, i)(\epsilon)\\ &\quad \leq \lambda(1-\beta)V_\beta^*(i)(\epsilon) + (1-\lambda)\bar{V}^*(i)(\epsilon). \qquad (25)\end{aligned}$$

Remark 5.1 By theorem 2.1 for all $i \in S$ and $\epsilon \in [0, \epsilon_0)$ we have: $\sup_{C(M)} \omega(\pi; i)(\epsilon) = \sup_C \omega(\pi; i)(\epsilon)$, where ϵ_0 is defined by definition 3.1. So we need only consider Markov policies. From now on, $C(M)$ will be the largest class of policies under consideration in this chapter.

Lemma 5.1 *Let $\Gamma(\beta, \lambda)$ be communicating or unichain, then there exists $\epsilon_3 > 0$ such that for $\epsilon \in [0, \epsilon_3)$ $\Gamma_\epsilon(\beta, \lambda)$ is communicating or unichain as well.*

Proof: Let us investigate the $d(j|i,a)(\epsilon)$ $i, j \in S$, $a \in A(i)$. Because of our assumption in this chapter (remark 3.1) and the finiteness of the state space and the action set, we have finite number of rational functions of ϵ. Let $\epsilon_3 = \min\{\epsilon_0, \min_{i,j,a}\{\epsilon(i,j,a) > 0 | \epsilon(i,j,a)$ is the smallest nonegative real root of function $d(j|i,a)(\epsilon) = 0\}\}$. By the definition of disturbance law and the communicating or unichain properties, the rest of the proof is trivial. □

Theorem 5.1 *Consider the process $\Gamma_\epsilon(\beta, \lambda)$, with $\lambda > 0$. Suppose $\Gamma(\beta, \lambda)$ is communicating or unichain. Then there exists $\epsilon_4 > 0$ such that for $\epsilon \in (0, \epsilon_4)$ we have: (i) For all $i \in S$: $\sup_C \omega(\pi; i)(\epsilon) = \lambda(1 - \beta)V_\beta^*(i)(\epsilon) + (1-\lambda)\bar{V}^*(i)(\epsilon)$; (ii) Given any $\delta > 0$, there exists $\pi(\delta) \in C(SUD)$ which is δ-optimal in*

$\Gamma_\epsilon(\beta,\lambda)$; *(iii) Let* $\delta > 0$ *and* $f^* \in C(D)$ *be optimal in* $\Gamma_\epsilon(A)$ *and* $g^* \in C(D)$ *be optimal in* $\Gamma_\epsilon(\beta)$, *where* $\epsilon \in (0,\epsilon_4)$. *There exists a non-random* $\tau(\delta,\lambda,\beta,\epsilon_4)$ *such that* $\pi^* = (f_0, f_1, \ldots)$ *with* $f_t = g^*$ *for* $t < \tau(\delta,\lambda,\beta,\epsilon_4)$ *and* $f_t = f^*$ *for* $t \geq \tau(\delta,\lambda,\beta,\epsilon_4)$, *is* δ*-optimal in* $\Gamma_\epsilon(\beta,\lambda)$ *for all* $\epsilon \in (0,\epsilon_4)$*; where* $\epsilon_4 = \min\{\epsilon_0,\epsilon_1,\epsilon_2,\epsilon_3\}$. *Moreover, (i), (ii) and (iii) hold also for the LWMDP* $\Gamma_0(\beta,\lambda)$.

Proof. Without loss of generality assume that $r(i,a) \geq 0$ for all $i \in S$ and $a \in A(i)$. We first prove (i) and (ii). For a fixed $\delta > 0$ and $\epsilon \in [0,\epsilon_1)$, by theorem 3.1, there exists a deterministic policy $f(\beta) \in C(D)$ and a positive integer $\tau(\delta,\lambda,\beta,\epsilon_1)$ such that for all $\tau \geq \tau(\delta,\lambda,\beta,\epsilon_1)$ we have for all $i \in S$:

$$\begin{aligned} V^*_\beta(i)(\epsilon) &= V_\beta(f(\beta);i)(\epsilon) \\ &= \sum_{t=0}^{\tau} \beta^t E_{f(\beta)}(R_t,i) + \sum_{t=\tau+1}^{\infty} \beta^t E_{f(\beta)}(R_t,i) \\ &\leq \sum_{t=0}^{\tau} \beta^t E_{f(\beta)}(R_t,i) + \frac{\delta}{\lambda(1-\beta)}. \qquad (26) \end{aligned}$$

By lemma 5.1, theorem 2.5 and theorem 3.2 for any $\epsilon \in (0,\min\{\epsilon_2,\epsilon_3\})$, there exists a deterministic policy $f(A) \in C(D)$ such that for all $i,j \in S$:

$$\begin{aligned} \bar{V}^*(i)(\epsilon) &= \bar{V}^*(j)(\epsilon) = \bar{V}(f(A);i)(\epsilon) \\ &= \bar{V}(f(A);j)(\epsilon) = \bar{V}(f(A)). \qquad (27) \end{aligned}$$

Hence $f(A)$ is also a maximizer in the process $\Gamma_0(A)$.

Define a policy $\pi = (f_0, f_1, \ldots) \in C(SUD)$ in the following way:

$$f_t = \begin{cases} f(\beta) & \text{if } t \leq \tau(\delta,\lambda,\beta,\epsilon_1); \\ f(A) & \text{if } t > \tau(\delta,\lambda,\beta,\epsilon_1). \end{cases} \qquad (28)$$

Then when $\epsilon \in [0,\min\{\epsilon_1,\epsilon_2,\epsilon_3\})$, we have $\lambda(1-\beta)V_\beta(\pi;i)(\epsilon) > \lambda(1-\beta)V^*_\beta(i)(\epsilon) - \delta$ for every $i \in S$. Note that by the communicating or unichain assumption and lemma 5.1, we have: $\omega(\pi;i)(\epsilon) > \lambda(1-\beta)V^*_\beta(i)(\epsilon) + (1-\lambda)\bar{V}^*(i)(\epsilon) - \delta$, for all $i \in S$ and $\epsilon \in [0,\epsilon_4)$ where $\epsilon_4 = \min\{\epsilon_0,\epsilon_1,\epsilon_2,\epsilon_3\}$. By (25) the proofs of (i) and (ii) are completed. By (i), (ii) and the proof of theorem 2.6 (see Krass [14]) (iii) is trivial. □

Corollary 5.1 *Let* $\Gamma_0(\beta,\lambda)$ *be unichain or communicating, then the theorem 5.1 remains valid.*

Proof. Because $\Gamma_0(\beta,\lambda)$ is unichain or communicating, we have that (27) holds for $\epsilon = 0$. By theorem 3.1 and theorem 3.2, theorem 5.1 is again valid. This proves the corollary. □

Remark 5.2 Corollary 5.1 is more general than theorem 5.1, but the conditions of theorem 5.1 are easy to check.

Without the communicating or unichain assumptions, we have :

Corollary 5.2 *Suppose that there is a policy* $\pi \in C(S)$ *such that* $\bar{V}(\pi;i) = \bar{V}^*(i) = v$ *for all* $i \in S$*, then the results (i), (ii) and (iii) of theorem 5.1 still apply.*

Proof. Investigate the proof of theorem 5.1 and note that we only need to establish (27). By the assumption of corollary, (27) is true. Hence, we can define π_t similarly to (28) for every t and then $\pi = (\pi_0, \pi_1, \ldots)$ is an δ-optimal policy. The proof is completed. □

When we relax the "scalar value" hypothesis on $\Gamma(\beta,\lambda)$ or on $\Gamma_0(\beta,\lambda)$, more difficulties arise. If we want to obtain similar results as in Krass [14] (the case without perturbation), we have to prove that theorem 2.7 is true uniformly for $\epsilon \in [0,\epsilon^*)$.

Theorem 5.2 *For the* $\Gamma_\epsilon(\beta,\lambda)$ *with* $\lambda > 0$ *and* $\epsilon \in [0,\epsilon_4)$*. Let* $f^* \in C(D)$ *be optimal in theorem 3.2. Given a fixed initial state* i *and an arbitrary* $\delta > 0$*, there exists an* $\epsilon_4 > 0$ *(see theorem 5.1) and a positive non-random* $\tau(\delta,\lambda,\beta,\epsilon_4)$ *such that for any* $\epsilon \in [0,\epsilon_4)$ *there exists a policy* $\pi(\epsilon) = (\pi_0(\epsilon), \pi_1(\epsilon), \ldots) \in C(SUD)$ *with* $\pi_t(\epsilon) = f^*$ *for all* $t \geq \tau(\delta,\lambda,\beta,\epsilon_4)$ *which is* δ*-*i*-optimal in* $\Gamma_\epsilon(\beta,\lambda)$.

Proof. Without loss of generality we assume that $r(i,a) \geq 0$ for all $a \in A(i)$ and $i \in S$. When $\beta = 0$, the result is trivial for the case of $\Gamma_\epsilon(A)$ by theorem 3.2. Hence we consider the case of $\beta > 0$. For the fixed $\delta > 0$ and $\epsilon \in [0,\epsilon_4)$, by remark 5.1, there exists a Markov policy $\bar{\pi}(\epsilon) = (\bar{\pi}_0(\epsilon), \bar{\pi}_1(\epsilon), \ldots)$ which is $\frac{\delta}{2}$-i-optimal in $\Gamma_\epsilon(\beta,\lambda)$. As in (26), there exists $\tau(\delta,\lambda,\beta,\epsilon_4)$ such that for all $\epsilon \in [0,\epsilon_4)$ and $\tau \geq \tau(\delta,\lambda,\beta,\epsilon_4)$ we have

$$\sum_{t=0}^{\tau} \beta^t E_{\bar{\pi}(\epsilon)}(R_t,i) \geq V_\beta(\bar{\pi}(\epsilon);i)(\epsilon) - \frac{\delta}{2\lambda(1-\beta)} \qquad (29)$$

uniformly for the $\epsilon \in [0,\epsilon_4)$ and for the fixed $i \in S$. Now for each $\bar{\pi}(\epsilon)$ we define $\pi(\epsilon) = (\pi_0(\epsilon), \pi_1(\epsilon), \ldots) \in C(SUD)$ with $\pi_t(\epsilon) = \bar{\pi}_t(\epsilon)$ for $t = 0, 1, \ldots, \tau(\delta,\lambda,\beta,\epsilon_4) - 1$ and $\pi_t(\epsilon) = f^*$ for $t \geq \tau(\delta,\lambda,\beta,\epsilon_4)$. Then similarly to the proof of theorem 2.7 (see Krass [14]), we obtain that for the fixed $i \in S$ and all $\epsilon \in [0,\epsilon_4)$ $\omega(\pi(\epsilon);i)(\epsilon) \geq \omega(\bar{\pi}(\epsilon);i)(\epsilon) - \frac{\delta}{2} \geq \sup_C \omega(\pi;i)(\epsilon) - \delta$.

This completes the proof of this theorem. □

Definition 5.1 *For any integer* τ *and a deterministic policy* $f \in C(D)$*, we define:*

$$C(\tau,f) = \{\pi = (\pi_0, \pi_1, \ldots) \in C(M) | \pi_t = f,\ t \geq \tau\}. \qquad (30)$$

The distance between two policies $\pi, \pi^* \in C(\tau,f)$ *is defined*

$$dis(\pi,\pi^*) = \max_{\substack{i \in S \\ a \in A(i) \\ t = 0,\ldots,\tau-1}} |\pi_t(a|i) - \pi^*(a|i)|. \qquad (31)$$

We say that $\pi \to \pi^*$ if and only if $dis(\pi, \pi^*) \to 0$, where $\pi, \pi^* \in C(\tau, f)$.

Lemma 5.2 *The class $C(\tau, f)$ is closed.*

Proof: Since the state, action spaces and τ are all finite, the proof is trivial. □

Lemma 5.3 *For any integer τ and a deterministic policy $f \in C(D)$, we have*

(i) For all $\pi \in C(\tau, f)$,

$$\omega(\pi; i)(0) = \underline{\lim}_{\epsilon\to 0} \omega(\pi; i)(\epsilon) = \lim_{\epsilon\to 0} \omega(\pi; i)(\epsilon). \qquad (32)$$

(ii) If $\pi \to \pi^*$, where $\pi, \pi^* \in C(\tau, f)$, then the twofold limit in (33), below, exists and we have:

$$\lim_{\substack{\epsilon \to 0 \\ \pi \to \pi^*}} \omega(\pi; i)(\epsilon) = \lim_{\pi\to\pi^*} \omega(\pi; i)(0)$$
$$= \lim_{\epsilon\to 0} \omega(\pi^*; i)(\epsilon) = \omega(\pi^*; i)(0). \qquad (33)$$

Proof: (i)

$$\begin{aligned}\omega(\pi; i)(\epsilon) & \\ &= \lambda(1-\beta)V_\beta(\pi; i)(\epsilon) + (1-\lambda)\bar{V}(\pi; i)(\epsilon) \\ &= \lambda(1-\beta)I(\epsilon) + (1-\lambda)II(\epsilon). \qquad (34)\end{aligned}$$

Because $\pi \in C(\tau, f)$, we have

$$\underline{\lim}_{\epsilon\to 0} I(\epsilon) = \overline{\lim}_{\epsilon\to 0} I(\epsilon) = V_\beta(\pi; i)(0). \qquad (35)$$

Similarly to the proof of Theorem 3.6.1 in Krass [14] and with the notation of (34), we have

$$\begin{aligned}
&\underline{\lim}_{\epsilon\to 0} II(\epsilon) \\
&= \underline{\lim}_{\epsilon\to 0} \bar{V}(\pi; i)(\epsilon) \\
&= \underline{\lim}_{\epsilon\to 0} \underline{\lim}_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} E_\pi(R_k, i)(\epsilon) \\
&= \underline{\lim}_{\epsilon\to 0} \underline{\lim}_{T\to\infty} \frac{1}{T} \left[\sum_{k=\tau}^{T-1} E_\pi(R_k, i)(\epsilon) + \sum_{k=0}^{\tau-1} E_\pi(R_k, i)(\epsilon)\right] \\
&= \underline{\lim}_{\epsilon\to 0} \underline{\lim}_{T\to\infty} \frac{1}{T} \sum_{k=\tau}^{T-1} E_\pi(R_k, i)(\epsilon) + 0 \\
&= \underline{\lim}_{\epsilon\to 0} \underline{\lim}_{T\to\infty} \frac{1}{T} \sum_{k=\tau}^{T-1} \sum_{i'\in S} \sum_{a\in A(i')} [\\
&\qquad P_\pi(X_k = i', Y_k = a | X_0 = i)(\epsilon) r(i', a)] \\
&= \underline{\lim}_{\epsilon\to 0} \underline{\lim}_{T\to\infty} \frac{1}{T} \sum_{k=\tau}^{T-1} \sum_{i'\in S} \sum_{a\in A(i')} \sum_{j\in S} [\\
&\qquad P_\pi(X_k = i', Y_k = a, X_\tau = j | X_0 = i)(\epsilon) r(i', a)] \\
&= \underline{\lim}_{\epsilon\to 0} \underline{\lim}_{T\to\infty} \frac{1}{T} \sum_{k=\tau}^{T-1} \sum_{i'\in S} \sum_{a\in A(i')} \sum_{j\in S} [\\
&\qquad P_\pi(X_k = i', Y_k = a | X_\tau = j, X_0 = i)(\epsilon) \times \\
&\qquad\quad P_\pi(X_\tau = j | X_0 = i)(\epsilon) r(i', a)] \\
&= \underline{\lim}_{\epsilon\to 0} \underline{\lim}_{T\to\infty} \frac{1}{T} \sum_{k=\tau}^{T-1} \sum_{i'\in S} \sum_{j\in S} [\\
&\qquad P_f(X_k = i', Y_k = f(i') | X_\tau = j)(\epsilon) \times \\
&\qquad\quad P_\pi(X_\tau = j | X_0 = i)(\epsilon) r(i', f(i'))] \\
&= \underline{\lim}_{\epsilon\to 0} \underline{\lim}_{T\to\infty} \frac{1}{T} \sum_{j\in S} P_\pi(X_\tau = j | X_0 = i)(\epsilon) \times \\
&\qquad \left[\sum_{k=\tau}^{T-1} \sum_{i'\in S} P_f(X_k = i', Y_k = f(i') | X_\tau = j)(\epsilon) \times \right. \\
&\qquad\quad \left. r(i', f(i'))\right] \\
&= \underline{\lim}_{\epsilon\to 0} \underline{\lim}_{T\to\infty} \frac{1}{T} \sum_{j\in S} P_\pi(X_\tau = j | X_0 = i)(\epsilon) \times \\
&\qquad\qquad \left[\sum_{k=0}^{T-\tau} E_f(R_k, j)(\epsilon)\right], \qquad (36)
\end{aligned}$$

where f is given in the lemma. Because of the finiteness of state space, action set and τ, we have

$$\begin{aligned}
&\sum_{j\in S} P_\pi(X_\tau = j | X_0 = i)(\epsilon) \underline{\lim}_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-\tau} E_f(R_k, j)(\epsilon) \\
&\quad \le \underline{\lim}_{T\to\infty} \frac{1}{T} \sum_{j\in S} P_\pi(X_\tau = j | X_0 = i)(\epsilon) \times \\
&\qquad\quad \sum_{k=0}^{T-\tau} E_f(R_k, j)(\epsilon) \\
&\quad \le \sum_{j\in S} P_\pi(X_\tau = j | X_0 = i)(\epsilon) \times \\
&\qquad\qquad \overline{\lim}_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-\tau} E_f(R_k, j)(\epsilon). \qquad (37)
\end{aligned}$$

Because $f \in C(D)$ and for any $j \in S$, we have

$$\begin{aligned}
\bar{V}(f; j)(\epsilon) &= \lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-\tau} E_f(R_k, j)(\epsilon) \\
&= \underline{\lim}_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-\tau} E_f(R_k, j)(\epsilon) \\
&= \overline{\lim}_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-\tau} E_f(R_k, j)(\epsilon) \\
&= [P^*(f)(\epsilon) r(f)](j). \qquad (38)
\end{aligned}$$

Hence, (36) is

$$\lim_{\epsilon\to 0} II(\epsilon) = \lim_{\epsilon\to 0}\sum_{j\in S} P_\pi(X_\tau = j|X_0 = i)(\epsilon)\times [P^*(f)(\epsilon)r(f)](j). \tag{39}$$

By lemma 3.1, continuity of rational functions of ϵ (see remark 3.1) and finiteness of S and $A(i)$ for $i\in S$, (39) is

$$\begin{aligned}
&\lim_{\epsilon\to 0} II(\epsilon)\\
&= \varliminf_{\epsilon\to 0}\sum_{j\in S} P_\pi(X_\tau = j|X_0 = i)(\epsilon)\,[P^*(f)(\epsilon)r(f)](j)\\
&= \varlimsup_{\epsilon\to 0}\sum_{j\in S} P_\pi(X_\tau = j|X_0 = i)(\epsilon)\,[P^*(f)(\epsilon)r(f)](j)\\
&= \lim_{\epsilon\to 0}\sum_{j\in S} P_\pi(X_\tau = j|X_0 = i)(\epsilon)\,[P^*(f)(\epsilon)r(f)](j)\\
&= \bar{V}(\pi;i)(0).
\end{aligned}\tag{40}$$

Combine (35) and (40) and (32) is proved.

(ii) Since the product of finitly many continuous functions is a continuous function, the finiteness of τ, S and $A(i)$ for $i\in S$, similarly to the proof of (i), yields (ii). □

Theorem 5.3 *Consider the $\Gamma_\epsilon(\beta,\lambda)$ with $\lambda > 0$. Let $f^*\in C(D)$ satisfy theorem 3.2. Given a fixed state and arbitrary $\delta > 0$, there exists $\epsilon^* > 0$, a positive non-random $\tau(\delta,\lambda,\beta,\epsilon^*)$ and a policy $\pi(\delta) = (f_0, f_1,\ldots)\in C(SUD)$ with $f_t = f^*$ for all $t\geq \tau(\delta,\lambda,\beta,\epsilon^*)$ such that $\pi(\delta)$ is δ-i-optimal in $\Gamma_\epsilon(\beta,\lambda)$ for all $\epsilon\in[0,\epsilon^*)$.*

Proof. Again we may assume that $r(i,a)\geq 0$ for all $a\in A(i)$ and $i\in S$. By theorem 5.2, for $\epsilon = 0$ we take $\pi(\delta)$ to be an $\frac{\delta}{2}$-i-optimal policy for LWMDP with switching time τ and $f^*\in C(D)$. Now assume that the statement of the theorem is wrong. Then there exists a sequence $\{\epsilon_n\}_{n=0}^{\infty}$, $\epsilon_n > 0$ such that (i) $\epsilon_n\to 0$; (ii) $\omega(\pi(\delta);i)(\epsilon_n) < \omega^*(i)(\epsilon_n) - \delta$. However by theorem 5.2, there exists an integer M such that when $n\geq M$, there exists a policy $\pi^n\in C(\tau, f^*)$ which is $\frac{\delta}{2}$-i-optimal, that is $|\omega(\pi^n;i)(\epsilon_n) - \omega^*(i)(\epsilon_n)|\leq\frac{\delta}{2}$; or $\omega(\pi^n;i)(\epsilon_n)\geq\omega^*(i)(\epsilon_n) - \frac{\delta}{2}$

By the finiteness of τ, S and $A(i)$ for $i\in S$, there exists a subsequence n_k of n such that $\pi^{n_k}\to\pi^*$. Without loss of generality we assume that the subsequence is the sequence $\{n\}$. By lemma 5.3,

$$\begin{aligned}
\omega(\pi^*;i)(0) &= \lim_{\substack{\epsilon\to 0\\ \pi\to\pi^*}}\omega(\pi;i)(\epsilon) = \lim_{n\to\infty}\omega(\pi^n;i)(\epsilon_n)\\
&\geq \lim_{n\to\infty}\omega^*(i)(\epsilon_n) - \frac{\delta}{2}\\
&\geq \lim_{n\to\infty}\omega(\pi(\delta);i)(\epsilon_n) + \delta - \frac{\delta}{2}\\
&= \omega(\pi(\delta);i)(0) + \frac{\delta}{2} > \omega^*(i)(0).
\end{aligned}\tag{41}$$

This is a contradiction to definition of $\omega^*(i)(0)$. □

References

[1] M. Abbad. *Perturbation and Stability Theory for Markov Control Problems.* PhD thesis, University of Maryland, USA, 1991.

[2] M. Abbad and J.A. Filar. Algorithms for Singularly Perturbed Markov Control Problems: A Survey. In C. Leondes, editor, *Control and Dynamic Systems.* Academic Press. (to appear).

[3] D. Blackwell. Discrete Dynamic Programming. *Ann. Math. Statist.*, 33:719–726, 1962.

[4] F. Delebecque. A Reduction Process for Perturbed Markov Chains. *SIAM J. of Applied Math.*, 48:325–350, 1983.

[5] C. Derman. *Finite State Markovian Decision Processes.* Academic Press, New York, 1970.

[6] C. Derman and R. Strauch. A Note on Memoryless Rules for Controlling Sequential Control Problems. *Ann. Math. Statist.*, 37:276–278, 1966.

[7] E.A. Feinberg and A. Shwartz. Markov Decision Models with Weighted Discounted Criteria. *Math. Oper. Res.*, 19:152–168, 1994.

[8] E. Fernandez-Gaucherand, M.K. Ghosh, and S.I. Marcus. Controlled Markov Processes on the Infinite Planning Horizon: Weighted and Overtaking Cost Criteria. *ZOR Methods and Models of OR*, 39:131–155, 1994.

[9] J.A. Filar and T.A. Schultz. Communicating MDPs: Equivalence and LP Properties. *Oper. Res. Letters*, 7:303–307, 1988.

[10] J.A. Filar and O.J. Vrieze. Weighted Reward Criteria in Competitive Markov Decision Processes. *ZOR Methods and Models of Operations Research*, 36:343–358, 1992.

[11] M.K. Ghosh and S.I. Marcus. Infinite horizon controlled diffusion problems with some nonstandard criteria. *J. Math. Sys. and Contr.*, 1:45–70, 1991.

[12] L.C.M. Kallenberg. *Linear Programming and Finite Markovian Control Problems*, volume 148 of *Mathematical Center Tracts.* Amsterdam, 1983.

[13] D. Krass, J.A. Filar, and S. Sinha. A Weighted Markov Decision Process. *Oper. Res.*, 40:465–470, 1992.

[14] Dmitry Krass. *Contributions to the Theory and Applications of Markov Decision Processes.* PhD thesis, The Johns Hopkins University, USA, July 1989.

[15] H. Mine and S. Osaki. *Markovian Decision Processes.* American Elsevier, New York, 1970.

[16] M.L. Puterman. Markov decision processes. In D.P. Heyman and M.J. Sobel, editors, *Handbooks in Operation Research and Management Science*, volume 2, pages 331–434, Amsterdam, 1990. North-Holland.

[17] S.M. Ross. *Introduction to Stochastic Dynamic Programming.* Academic Press, New York, 1983.