Policy Improvement Algorithm for Singularly Perturbed Discounted Markov Decision Processes

Mohammed Abbad [1] *Abdesselam El bahja* [2]

[1] Département de Mathématiques et Informatiques
Facult des sciences de Rabat
B.P: 1014, Morocco

[2] Ecole des Mines de Rabat /Agdal
Département informatique
B.P: 753, Morocco

Abstract

In this paper, we consider a perturbed Markov decision process with the discounted reward criterion .The transition probabilities and discount factor are perturbed slightly.We assume that the underlying process is completely decomposable in finite number of separate irreducible processes .We introduce the limit Markov control problem which is the optimization problem that should be solved in case of singular perturbations. In order to solve the limit Markov control problem, we propose an aggregation-disaggregation policy improvement algorithm which converges in a finite number of iterations to an optimal deterministic strategy.

1 Introduction

Finite state and action Markov decision processes (MDPs for short) are dynamic, stochastic, systems controlled by some controller, sometimes referred to as “decision make”. These models have been extensively studied since 1950’s by applied probabilists, operations researchers, and engineers.

Engineers typically refer to these models as “Markov control problems”, and in this paper we shall use these labels interchangeably. The early MDP models were studied by Howard [21] and Blackwell [9] and, following the latter, are sometimes referred to as “Discrete Dynamic Programming”.

During the 1960's and 1970's the theory of classical MDP's evolved to the extent that there is now a complete existence theory, and a number of good algorithms for computing optimal policies, with respect to criteria such as maximization of limiting average expected output, or the discounted expected output (eg. see [7], [13], [17], [20], [23]). These models were applied in a variety of contexts, ranging from water-ressource models, through communication networks, to inventory and maintenance models.

One class of problems that began to be addressed in recent years focussed around the following question:

How is the analysis of an MDP model affected by perturbations (typically small) of the problem data?

If the perturbation of a Markov chain alters the ergodic structure of that chain, then stationary distribution of the perturbed processus has a discontinuity at the zero value of the disturbance parameter.

This phenomenon was illustrated by Schweitzer [18] with the following example:

Let $P_\epsilon = \begin{bmatrix} 1-\epsilon/2 & \epsilon/2 \\ \epsilon/2 & 1-\epsilon/2 \end{bmatrix}$ be the perturbed Markov chain whose stationary distribution matrix is:

$P_\epsilon{}^* = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$ for all $\epsilon \in [0,2]$. Thus we have:

$\underset{\epsilon\to 0}{Lim}\ P_\epsilon{}^* = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \neq P_0{}^* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, where $P_0{}^*$ is the stationary distribution matrix of the unperturbed Markov chain P_0.

Some authors [1], [2], [3], [5], [8] considered a singularly perturbed MDP with the limiting average reward criterion.

In this paper we consider a singular perturbation with the discounted expected criterion. We give explicitly the limit Markov control problem (limit MCP) that is entirely different from the original unperturbed MDP, which forms an appropriate asymptotic to a whole family of perturbed problems; thus only the single limit MCP needs to be solved . We construct an aggregation-disaggregation algorithm for solving the limit MCP, which is the main contribution of this paper.

2 Definitions and Preliminaries

A discrete Markovian decision process (MDP, for short) is observed at discrete time points $t = 0, 1, 2, ...$ The state space is denoted by $S = \{1, 2, ..., N\}$. With each state $s \in S$ we associate a finite action set $A(s) = \{1, 2, ..., m(s)\}$. At any time point t, the process is in one of the states s and the controller chooses an action $a \in A(s)$; as a result the following occur : i) an immediate reward $r(s,a)$ is accrued , and ii) the process moves to state $s' \in S$ with transition probability

$$p(s'/s,a), \text{ where } p(s'/s,a) \geq 0 \text{ and } \sum_{s' \in S} p(s'/s,a) = 1.$$

Henceforth, such an MDP will be synonymous with the four-uple:
$\Gamma = < S; [A(s), s \in S]; [r(s,a), s \in S, a \in A(s)]; [p(s'/s,a),\ s, s' \in S, a \in A(s) >$.

While a general control state in Γ may depend on the complete state-action histories of the process, in this paper we shall concern ourselves only with the class F_S of all stationary strategies . A stationary strategy $\pi \in F_S$ is the vector: $\pi = (\pi(s,a)/(s,a) \in S \times A(s))$ where $\pi(s,a)$ is the probability that controller chooses action $a \in A(s)$ in state s whenever that state is visited; of course; $\sum_{a \in A(s)} \pi(s,a) = 1$ for all s.

A strategy $\pi \in F_S$ will be called deterministic if $\pi(s,a) \in \{0,1\}$ for all $(s,a) \in S \times A(s)$.

With every $\pi \in F_S$ we associate the following quantities : $r(\pi) = (r_1(\pi), ..., r_N(\pi))^T$; the vector of single stage expected rewards in which $r_s(\pi) := \sum_{a \in A(s)} r(s,a)\pi(s,a)$ for each $s \in S$; a Markov matrix $P(\pi) = (P_{ss'}(\pi)^N_{s,s'=1})$; where $P_{ss'}(\pi) := \sum_{a \in A(s)} P(s'/s,a)$ $\pi(s,a)$ for all $s, s' \in S$; the generator of the corresponding Markov chain, namely, the matrix $G(\pi) := P(\pi) - I$; the corresponding Cesaro-limit matrix is defined by:

$P^*(\pi) := (P^*_{ss'}(\pi)^n_{s,s'=1}) := \lim_{t \to +\infty} 1/t + 1 \ \sum_{k=0}^{t} P^k(\pi)$, where $P^0(\pi) := I_N$, an $N \times N$ identity matrix.

The classical discount expected Markov decision problem is the optimisation problem:

$$\text{Find } \pi^* \in F_S \text{ such that: } V_\beta(s,\pi^*) \geq V_\beta(s,\pi)\ \forall \pi \in F, \forall s \in S \tag{1.1}$$

where; $V_\beta(s,\pi) = \sum_{t=0}^{+\infty} \beta^t E_{s\pi}(R_t)$; $\beta \in]0,1[$ is the discounted factor, R_t is the randon variable which represents the reward at time t, and $E_{s\pi}(R_t)$ is the expected reward at time t when the process begins at state s and the controller uses the strategy π. The following theorem is well known (eg. see [15], [22]).

Theorem(1.1) $V^*_\beta = \max_{\pi \in F} V_\beta(\pi)$ exists and $V^*_\beta = \max_{\pi \in F_S} [r(\pi) + \beta P(\pi) V^*_\beta]$ where

$V_\beta(\pi) = V_\beta(s,\pi)^T_{s\in S}$.

Remark (1.1)

Let $\beta = 1/1+\lambda, \lambda > 0$; and $V_\lambda(\pi) = (1-\beta)V_\beta(\pi)$. Then:

i) $V_\lambda(\pi) = \lambda(1+\lambda)^{-1}\sum_{m=0}^{+\infty}(1+\lambda)^{-m}P^m(\pi)r(\pi) = \lambda(1+\lambda)^{-1}[I_N-(1+\lambda)^{-1}P(\pi)]^{-1}r(\pi)$

ii) $V^*_\lambda := \max_{\pi\in F_S} V_\lambda(\pi) = (1-\beta)\max_{\pi\in F_S} V_\beta(\pi)$

iii) $\underset{\lambda\to 0}{Lim}\ V_\lambda(\pi) = P^*(\pi)r(\pi)$.

From theorem (1.1) and part (ii) of remark 1.1, we can get the following result:

Proposition (1.1) V^*_λ satisfies the optimality equation:

$$-\lambda V^*_\lambda + \max_{\pi\in F_S}\{G(\pi)V^*_\lambda + \lambda r(\pi)\} = 0.$$

A strategy π^* satisfying the equation above will be called optimal. It is well known that there always exist an optimal deterministic strategy and there is a number of finite algorithms for its computation (e.g; [15], [22]). In this paper, we shall assume the following:

A_1) $S = \bigcup_{i=1}^{n} S_i$, where $S_i \cap S_j = \emptyset$ if $i \neq j, n > 1$,
card $S_i = n_i, n_1 + n_2 + ... + n_n = N$

A_2) $p(s'/s,a) = 0$, whenever $s \in S_i,\ s' \in S_j,\ i \neq j$ and $a \in A(s)$

Consequently we can think of Γ as being the union of n smaller MDP's Γ_i , defined on the state space S_i for each $i = 1, 2, ...n$, respectively. Note that if F_i is the space of stationary strategies in Γ_i, then a strategy $\pi \in F$ in Γ can be written in the natural way as $\pi = (\pi^1, \pi^2, ..., \pi^n)$ where $\pi^i \in F_i$.

The probability transition matrix in Γ_i corresponding to π^i is of course defined by:

$P_i(\pi^i) = (p_{ss'}(\pi^i))_{s,s'\in S_i}$; and the generator $G_i(\pi^i)$ and the Cesaro-limit matrix ${P_i}^*(\pi^i)$ can be defined in a manner analogous to that in the original process Γ.

In addition, we assume the following:

A_3) For every $i = 1, ..., n$ and for all strategy $\pi^i \in F_i$, the matrix $P_i(\pi^i)$ is an irreducible matrix.

In view of $A_3)$ $P_i{}^*(\pi^i)$ is a matrix with identical rows. We shall denote any row of $P_i{}^*(\pi^i)$ by $p_i^*(\pi^i)$.

Remark (1.2)
Note that for all $\pi \in F_s$ we have the following representation of $P^*(\pi) : P^*(\pi) = EM(\pi)$ where E is an $N \times n$ matrix with entries:

$$e_{sj} = \begin{cases} 1 \quad if \ \sum_{k=1}^{j-1} n_k < s \leq \sum_{k=1}^{j} n_k \\ \\ 0 \quad otherwise \end{cases}$$

for $s = 1, 2, ..., N$ and $j = 1, 2, ...n$; and $M(f)$ is an $n \times N$ matrix with entries:

$$m_{js}(\pi) = \begin{cases} [p_j^*(\pi^j)]_s \quad if \ \sum_{k=1}^{j-1} n_k < s \leq \sum_{k=1}^{j} n_k \\ \\ 0 \qquad otherwise \end{cases} \qquad \text{of}\, j = 1, 2, ...n \text{ and } s = 1, 2, ..., N.$$

Of course we set $\sum_{k=1}^{0} n_k := 0$. Note also that from the above definitions we conclude that ; $M(\pi)E = I_n$.

3 Perturbations

In order to analyse the perturbed Markov control problem, we must first understand the uncontrolled case that is equivalent to the controller having only a single strategy at his disposal. This sub-topic is sometimes called the perturbation theory of Markov chains (M.Cs.) and is of interest in its own right.

In what follows, we concentrate only on a specially structured case that has received the most attention in the litterature. A nearly decomposable M.C is defined by an $N \times N$ irreducible transition probability matrix $P(\epsilon)$,

$$P(\epsilon) = P + \epsilon D; \text{where}$$

$$P = \begin{bmatrix} P_1 \ \ 0 .... \ 0 \\ 0 \ \ P_2 \ 0 .... \ 0 \\ . \\ . \\ 0 \ \ 0 \ .... P_n \end{bmatrix}$$

and P_i is an $n_i \times n_i$ irreducible transition probability matrix $i = 1, 2, ..., n$.

This class arises naturally in many applications of large scale finite state Markov chains. It is characterized by a decomposition of the states into groups, with strong interactions between states in the same groups, and weak interactions between states in different groups.

The strong-weak interaction structure was first introduced by Simon and Ando [24]. Courtois [10] developed the first analytical techniques for this class, and applied it to many problems in queueing networks and computer systems.

The fundamental problem to be analyzed for this class of M.Cs is the computation of the stationary distribution .This problem suffers from high dimensionality and ill conditioning. Courtois gave an aggregation procedure for the computation of an $0(\epsilon)$-approximation of the stationary distribution of $P(\epsilon)$. But it can be easily checked that Courtois's procedure will fail in many cases where the nearly decomposable structure is no longer present. This points to the need to develop an analogous theory for the more general perturbed M.Cs.

Based on the theory of Kato [16] for the perturbation of linear operators, Delebecque [11] derived a more general formula for the approximation of the stationary distribution matrix.

In this section we introduce the formulation and some results of the underlying control problem for the singularly perturbed MDP; the so called "limit Markov control problem"(Limit MCP).In particular, we prove that an optimal solution to the perturbed MDP can be approximated by an optimal solution of the limit MCP for sufficiently small perturbation.

We shall now consider the situation where the transition probabilities of Γ are perturbed slightly.

Towards this goal we shall define the disturbance law as the set:

$D = \{d(s'/s, a)/(s, a, s') \in S \times A(s) \times S\}$ where the elements of D satisfy:

$\sum_{s' \in S} d(s'/s, a) = 0$ (for all $(s, a) \in S \times A(s)) - 1 \leq d(s/s, a) \leq 0, d(s'/s, a) \geq 0, s \neq s'; (s, a, s') \in S \times A(s) \times S$.

Now, with every $\pi \times F_s$ we can associate a perturbation generator matrix $D(\pi) = [d_{ss'}(\pi)]_{s,s' \in S}$ where; $d_{ss'}(\pi) = \sum_{a \in A(s)} d(s'/s, a)\pi(s, a)$ and we shall also require that there exists $\epsilon_0 > 0$ such that for every $\pi \in F_s$:

$G_\epsilon(\pi) = G(\pi) + \epsilon D(\pi)$ is a generator of a Markov chain for any $0 \le \epsilon \le \epsilon_0$.

We shall consider a family of perturbed processes Γ_ϵ for $0 \le \epsilon \le \epsilon_0$ that differ from the original MDP Γ only in the transition law, namely in Γ_ϵ for every $s, s' \in S$, and $a \in A(s)$ we have that: $p_\epsilon(s'/s,a) = p(s'/s,a) + \epsilon d(s'/s,a)$.

We have that every $\pi \in F_s$ induces in the perturbed process Γ_ϵ the Markov chain with the probability transition matrix $P_\epsilon(\pi) = G_\epsilon(\pi) + I_N$.

(SPA) Singular perturbation assumption:

For every $\pi \in F_S$ and $\epsilon \in (0, \epsilon_0]$, $P_\epsilon(\pi)$ is an irreducible matrix.

Under (SPA), the MDP Γ_ϵ, $\epsilon \in (0, \epsilon_0]$, defined as:

$\Gamma_\epsilon =< S; [A(s), s \in S]; [r(s,a), s \in S, a \in A(s)]; [p_\epsilon(s'/s,a), s, s' \in S, a \in A(s)] >$ is called the singular perturbed MDP.

As in Delebecque and Quadrat [12], we shall also perturb the discounted factor in the following manner:

$\lambda = \epsilon\mu \; ; \epsilon > 0$ and $\mu > 0$

Now, in the perturbed discounted problem Γ_ϵ, for every stationary strategy π, we have that:

$$V_\lambda(\pi) = V_{\mu\epsilon}(\pi) = \mu\epsilon(1+\mu\epsilon)^{-1} \sum_{m=0}^{+\infty} (1+\mu\epsilon)^{-m} {P_\epsilon}^m(\pi) r(\pi).$$

Set $V_\epsilon(\pi) = V_{\mu\epsilon}(\pi)$, then:

$$V_\epsilon(\pi) = \mu\epsilon(1+\mu\epsilon)^{-1} \sum_{m=0}^{+\infty} (1+\mu\epsilon)^{-m} {P_\epsilon}^m(\pi) r(\pi).$$

Let ${V_\epsilon}^*(s) = \max_{\pi \in F_S} V_\epsilon(s,\pi), s \in S$.

The optimality equation in proposition (1.1) becomes:

$$-\mu\epsilon {V_\epsilon}^*(s) + \max_{\pi \in F_S}\{[G_\epsilon(\pi).{V_\epsilon}^*](s) + \epsilon\mu r(s,\pi)\} = 0 \text{ for all } s \in S \qquad [2.3]$$

In Phillips and Kokotovic [19], the authors came up with equation [2.3]. They proposed an algorithm for solving equation [2.3].

In [19], the authors considered the continuous M.C model:

$$dp/d\tau = p(G + \epsilon D) \tag{1}$$

where $I + G + \epsilon D$ is a nearly decomposable M.C, and p is the N-dimensional row vector whose entries are the probabilities p_i of being in state i at time τ.

In order to analyse the influence of weak interactions ϵD, the authors considered the change time scale to $t := \epsilon\tau$. Therefore, in the t-scale the model (1) becomes:

$$dp/dt = p(G/\epsilon + D) \tag{2}$$

In the discrete time, the model (2) has the analog:

$$p(k+1) = p(k)(G/\epsilon + I + D) \tag{3}$$

It is well known (eg.see [22]) that the optimality equation with respect to the discounted reward criterion for the model decribed above is:

$$J_\epsilon{}^\alpha = \max_{\pi\in F_D}\{\alpha(G(\pi)/\epsilon + D(\pi) + I)J_\epsilon{}^\alpha + r(\pi)\}. \tag{4}$$

Note that (4) is similar to [2.3]. Also, Delebecque and Quadrat [12] came up with equation [2.3] and they proposed another algorithm.

In this paper, our main objective is to solve equation [2.3] for small ϵ, by using the methods developed in [5], [8]. Using similar techniques, Abbad and Kissai [4] found an algorithm for solving [2.3] which is based on linear programming.

For each $\pi \in F_S$ let us define the $n \times n$ matrix $\overline{B}(\pi)$ by:

$$\overline{B}(\pi) = M(\pi)D(\pi)E \text{ and } \overline{P}(\pi) = \overline{B}(\pi) + I_n$$

Note that $\overline{B}(\pi)$ is a generator of an aggregated M.C on state space $\overline{S} = \{1, 2, ..., n\}$ with the transition probability matrix $\overline{P}(\pi)$.

Definition 2.1

For every $\pi \in F_s$ we define:

$$\overline{V}(i,\pi) = \mu(1+\mu)^{-1}\{[I - (1+\mu)^{-1}\overline{P}(\pi)]^{-1}\overline{r}(\pi)\}_i, i \in \overline{S} \ ,$$

$$\overline{r}(\pi) = M(\pi)r(\pi)$$

$$\hat{V}(s,\pi) = E\,\overline{V}(i,\pi) \text{ for all } s \in S_i$$

Proposition 2.1

For all strategy $\pi \in F_s$, we have: $V_\epsilon(\pi) = \hat{V}(\pi) + 0(\epsilon)$

Proof: (see [12]).

Remark 2.2

$\underset{\epsilon \to 0}{Lim}\, V_\epsilon(s, \pi) = \hat{V}(s, \pi) = V(i, \pi)$; for all $s \in S_i$, and $\pi \in F_s$

4 Limit Markov Control Problem

Definition 3.1

The optimization problem: $V^*(s) = \max_{\pi \in F_S} \hat{V}(s, \pi); s \in S$ [L] is called the limit Markov control problem.

The problem:

$\overline{V}^*(i) = \max_{\pi \in F_S} \overline{V}(i, \pi) \quad i \in \{1, 2, ..., n\}$ [AL]

is called the aggregated limit Markov problem.

Remark 3.1

We have that $\hat{V}(s, \pi) = \overline{V}(i, \pi)$ for all $s \in S_i, i = 1, 2, ..., n$ and $\pi \in F_s$;

thus $\hat{V}^*(s) = \overline{V}^*(i), s \in S_i \quad i = 1, 2, ..., n$.

It follows that any optimal strategy for [AL] is also an optimal strategy for [L] and vice- versa.

Proposition 3.1

$V_\epsilon{}^* = \hat{V}^* + 0(\epsilon)$

Proof: (see [6]).

Remark 3.2

By proposition 3.1 we have that:

$\underset{\epsilon \to 0}{Lim}\, V_\epsilon{}^*(s) = \hat{V}^*(s)$, for all $s \in S$.

Proposition 3.2

There exists a deterministic strategy $f \in F_D$ such that:

$\hat{V}(s,\pi) \leq \hat{V}(s,f)$, for all $s \in S$, and $\pi \in F_s$.

Proof :

Let $(\epsilon_n)_{n=1}^{+\infty}$ be any sequence in $(0,\epsilon_0]$ which converges to 0. From Markov decision theory we have that for any n there exists $f\epsilon_n \in F_D$ such that:

$V\epsilon_n(s,\pi) \leq V\epsilon_n(s,f\epsilon_n)$ for each $s \in S, \pi \in F_s$.

Since $f\epsilon_n \in F_D$, and F_D is finite, there must exist a deterministic strategy $f \in F_D$ and subsequence ${(\epsilon_{n_k})_{k=1}}^{\infty}$ of the sequence ${(\epsilon_n)_{n=1}}^{\infty}$ such that:

$V\epsilon_{n_k}(s,\pi) \leq V\epsilon_{n_k}(s,f)$ for each $k \in \mathbf{N}^*$, $s \in S$, and $\pi \in F_s$.

From proposition 2.1, it follows that for all $s \in S, \pi \in F_s$:

$$\begin{aligned}\hat{V}(s,\pi) &= \underset{\epsilon\to 0}{Lim}\, V_\epsilon(s,\pi)\\ &= \underset{k\to+\infty}{Lim}\quad V\epsilon_{n_k}(s,\pi).\end{aligned}$$

Therefore $\hat{V}(s,\pi) \leq \hat{V}(s,f)$, for each $s \in S$, $\pi \in F_s$ □

In view of proposition (3.2) we conclude that the problem [L] can be restricted to the optimization problem [L']:

$$\max_{\pi \in F_D} \hat{V}(\pi);$$

and any optimal strategy for [L'] is also optimal for [L].

Theorem 3.1

There exists a deterministic strategy $f^\circ \in F_D$ and a number $\delta > 0$ such that for all $\epsilon \in]0,\delta[$, f° is optimal for Γ_ϵ. Moreover f° is optimal for [L].

Proof:

From Markov decision theory, for any $\epsilon \in]0,\epsilon_0[$, there exists an optimal deterministic strategy $f^\circ \in F_D$ for the problem (Γ_ϵ).

Since the class F_D is finite there exists a deterministic strategy f° and a sequence ${(\epsilon_n)_{n=1}}^{\infty}$ in $]0,\epsilon_0[$ which converges to 0 such that f° is an optimal strategy in $[\Gamma\epsilon_n]$

for all $n \in \mathbf{N}^*$.

For a fixed f in F_D and $s \in S$, we have that:

$V\epsilon_n(s, f^\circ) \geq V\epsilon_n(s, f); n \in \mathbf{N}^*$.

From the fact that $V_\epsilon(s, f^\circ)$ and $V_\epsilon(s, f)$ are rational functions of ϵ; there exists $\epsilon(s, f)$ in $]0, \epsilon_0[$ such:

For all ϵ in $]0, \epsilon(s, f)[$; $V_\epsilon(s, f^\circ) \geq V\epsilon(s, f)$.

Define $\delta := \min\{\epsilon(s, f), s \in S, f \in F_D\} \in]0, \epsilon_0[$. Now we have that $V_\epsilon(s, f^\circ) \geq V_\epsilon(s, f)$ for any $\epsilon \in]0, \delta[$, $s \in S$ and $f \in F_D$.

This proves the first part of the theorem.

For the second part let $\epsilon \to 0$, then Remark 2.2 implies that:

$\hat{V}(f^\circ) \geq \hat{V}(f)$ for all f in F_D. □

Corollary 3.1 (limit control principle)

Let $\pi^\circ \in F_D$ be any optimal strategy in [L], then for all $\mathcal{B} > 0$ there exists $\epsilon_{\mathcal{B}}$ such that for all $\epsilon \in]0, \epsilon_{\mathcal{B}}[: |V_\epsilon{}^*(s) - V\epsilon(s, \pi^\circ)| < \mathcal{B}$ for all $s \in S$.

Proof:

Let ϵ be any number in $]0, \delta[$, in view of theorem 3.1, for all s in S we have that:

$|V_\epsilon(s, \pi^\circ) - V_\epsilon{}^*(s)| = |V_\epsilon(s, \pi^\circ) - \hat{V}(s, \pi^\circ) + \hat{V}(s, f^\circ) - V_\epsilon{}^*(s)|$

where f° is as in theorem 3.1 . Since $V_\epsilon{}^* = V_\epsilon(f^\circ)$, we shall write:

$|V_\epsilon(s, \pi^\circ) - V_\epsilon{}^*(s)| \leq (V_\epsilon(s, \pi^\circ) - \hat{V}(s, \pi^\circ)| + |\hat{V}(s, f^\circ) - V_\epsilon(s, f^\circ)|$

In view of $\underset{\epsilon \to 0}{Lim}\, V_\epsilon = \hat{V}$; then for all $\mathcal{B} > 0$ there exists $\epsilon_{\mathcal{B}}$ such that:

$|V_\epsilon(s, \pi^\circ) - \hat{V}(s, \pi^\circ)| < \mathcal{B}/2$ and $|\hat{V}(s, f^\circ) - V_\epsilon(s, f^\circ)| < \mathcal{B}/2$ for all s in S.

The next result shows that remark 3.2 can be proved without using proposition 3.1.

Theorem 3.2

$\underset{\epsilon \to 0}{Lim}\, V_\epsilon{}^* = \hat{V}{}^*$.

Proof:

$$\begin{aligned}\hat{V}^{\,*} &= \max_{f\in F_S}\hat{V}(f) \;\text{ (by definition 2.3.1)}\\ &= \hat{V}(f^{\circ}) \;\text{ (by theorem 3.1)}\\ &= \mathop{Lim}_{\epsilon\to 0} V_\epsilon(f^{\circ}) \;\text{ (by remark 2.2)}\\ &= \mathop{Lim}_{\epsilon\to 0} V_\epsilon{}^{*} \;\text{ (by theorem 3.1).} \quad \square\end{aligned}$$

5 Aggregated Problem and Policy Improvement Algorithm

In section 3, we proved that the limit M.C.P (L): $\max_{\pi\in F_s}\hat{V}(\pi)$ where $\hat{V}(\pi) = \mathop{Lim}_{\epsilon\to 0} V_\epsilon(\pi)$ can be converted to an equivalent aggregated problem:

$$\max_{\pi\in F_S}\overline{V}(\pi) \quad \text{where} \quad \overline{V}(\pi) = \mu/(1+\mu)\left[I_n - \frac{1}{1+\mu}\overline{P}(\pi)\right]^{-1}\overline{r}(\pi). \tag{4.0}$$

the vector $\overline{V}(\pi)$ in (4.0) can be considered as the reward of the strategy π in some M.C.P $\overline{\Gamma}$ that we shall define as follows:

1) the state space of $\overline{\Gamma}$ is $\overline{S} = \{1, 2, 3, \ldots\ldots., n\}, (i \equiv S_i)$,

2) the action space of $\overline{\Gamma}$ is $\overline{A}(i) := \prod_{s\in S_i} A(s)$ for each $i \in \overline{S}$,

3) the transition law of $\overline{\Gamma}$ is : for all $i, j \in \overline{S}; a \in \overline{A}(i)$;

$$q(j/i, a) = \begin{cases} 1 + \sum_{s'\in S_i}\sum_{s\in S_i}(p_i{}^{*}(a))_s\; d(s'/s, a_s) & i = j \\ \\ \sum_{s'\in S_j}\sum_{s\in S_i}(p_i{}^{*}(a))_s\; d(s'/s, a_s) & i \neq j \end{cases}$$

4) the rewards for $\overline{\Gamma}$: for all $i \in \overline{S}, a \in \overline{A}(i)$, where $a = (a_s/s \in S_i)$

$$c(i, a) = \sum_{s\in S_i} (p_i^{*}\ (a))_s r(s, a_s).$$

Remark 4.1

For all $i \in \overline{S}$, action $a \in \overline{A}(i)$ defines a deterministic strategy in Γ_i, which takes action a_s in state $s \in S_i$. If $\overline{\pi}$ is a deterministic strategy in $\overline{\Gamma}$ and π its corresponding deterministic strategy in Γ, then $\overline{V}(\pi) = \dfrac{\mu}{1+\mu}\,\overline{V}_\mu(\overline{\pi})$, where $\overline{V}_\mu(\overline{\pi})$ is the reward for the strategy $\overline{\pi}$ in the aggregated problem $\overline{\Gamma}$ with the discounted criterion (discounted factor is $\frac{1}{1+\mu}$).

Since problems (AL) and ($\overline{\mathrm{L}}$) are equivalent, we shall solve the problem

$$(\overline{\mathrm{L}}) \; : \; \max_{\overline{\pi}} \overline{V}_{\mu}(\overline{\pi})$$

We know that the problem ($\overline{\mathrm{L}}$) admits an optimal solution (proposition 3.2 and remark 3.1). From remark 4.1, it follows that the problem ($\overline{\mathrm{L}}$) can be solved by using the policy improvement algorithm (eg.see [15]):

Algorithm 1

step 1: Select an arbitrary deterministic strategy $\overline{\pi}$ and compute:

$$\overline{V}_{\mu}(\overline{\pi}) = \left[I_n - \frac{1}{1+\mu} Q(\overline{\pi}) \right]^{-1} C(\overline{\pi}) \quad , \tag{4.1}$$

where $C(\overline{\pi}) = (c(i, \overline{\pi}(i))_{i \in \overline{S}}$.

step 2: For all $i \in \{1, 2, ...n\} = \overline{S}$, find $a \in \overline{A}(i)$ that satisfies:

$$\left\{ c(i,a) + \beta \sum_{j=1}^{n} q(j/i,a) \overline{V}_{\mu}(j, \overline{\pi}) \right\} > \overline{V}_{\mu}(i, \overline{\pi}) \; ;$$

where $\beta = \frac{1}{1+\mu}$ and $\overline{V}_{\mu}(\overline{\pi}) = (\overline{V}_{\mu}(i, \overline{\pi}))_{i \in \overline{S}}$.

step 3: Let $\overline{\pi}_1$ be the deterministic strategy defined by: for all $i \in \overline{S}$,

$$\overline{\pi}_1(i) = \begin{cases} a & if \; a \; exists \\ \overline{\pi}(i) & if \; a \; does \; not \; exist \; . \end{cases}$$

step 4: If $\overline{\pi}_1 = \overline{\pi}$, then $\overline{\pi}$ is optimal (stop).

step 5: $\overline{\pi}_1(i) \; \rightarrow \; \overline{\pi}(i), i = 1, 2, ..., n$ and go to step 1.

We shall now develop the fundamental steps of algorithm 1.

Step1.

We select an arbitrary deterministic strategy $\overline{\pi} \in \overline{\Gamma}$ and compute:

${}^*q_{ij}(\overline{\pi}) \; := \; q(j/i, \overline{\pi}(i))$ for $i, j \; \in \; \overline{S}$ and $Q(\overline{\pi}) \; = \; (q_{ij}(\overline{\pi}))_{i,j=1,......n}$ for $i, j \; \in \overline{S}$. ${}^*C_i(\overline{\pi}) := c(i, \overline{\pi}(i)), i \in \overline{S}$

${}^*C(\overline{\pi}) := (C_1(\overline{\pi}), C_2(\overline{\pi}), ..., C_n(\overline{\pi}))$.

Remark 4.2

If $\overline{\pi}$ is a deterministic strategy in $\overline{\Gamma}$, then the corresponding deterministic strategy π in Γ is defined by:

$\pi(s) = [\overline{\pi}(i)]_s, s \in S_i$ and $i \in \overline{S}$

Remark 4.3

In $q_{ij}(\overline{\pi})$ and $C_i(\overline{\pi})$ we must compute $(p_i{}^*(\overline{\pi}(i))$.

For computing $(p_i{}^*(\overline{\pi}(i))$; $i \in \overline{S}$; we can apply the following algorithm [eg. see [15]] in which $P_i{}^*(\overline{\pi}(i))$ is irreducible for each $i \in \overline{S}$.

Algorithm 2 (for computing $p_i^*(\overline{\pi}(i)), i \in \overline{S}$).

1- Solve the (**steady-state equations**)

$$x_s{}^i = \sum_{s' \in S_i} x_{s'}{}^i p_{s's}, s \in S \ . \quad p_{s's} = p(s'/s, [\overline{\pi}(i)]_s)$$

$$\sum_{s \in S_i} x_s{}^i = 1$$

2- Vector $p_i^*(\overline{\pi}(i))$ is given by $[p_i^*(\overline{\pi}(i))]_s = x_s{}^i, \ s \in S_i$.

Then compute $[I_n - \dfrac{1}{1+\mu} Q(\overline{\pi})]^{-1} C(\overline{\pi})$.

Step 2

Now, we shall show that for each $i = 1, 2, ..., n$ the problem in step2 can be converted to one iteration of the policy improvement algorithm for some MDP defined by Γ_i, except for the rewards which will be defined appropriately.

For every $i \in \overline{S}$ and $a \in \overline{A}(i)$ we have that:

$$\{c(i,a) + \beta \sum_{j=1}^{n} q(j/i,a) \overline{V}_\mu(j,\overline{\pi})\} = \sum_{s \in S_i} (p_i{}^*(a))_s r(s, a_s)$$

$$+\beta\{\sum_{\substack{j=1 \\ i \neq j}}^{n} [\sum_{s' \in S_j} \sum_{s \in S_i} (p_i{}^*(a))_s d(s'/s, a_s) \overline{V}_\mu(j,\overline{\pi}))]$$

$$+(1 + \sum_{s' \in S_i} \sum_{s \in S_i} (p_i{}^*(a))_s d(s'/s, a_s)) \overline{V}_\mu(i,\overline{\pi})\}$$

$$= \beta\overline{V}_\mu(i,\overline{\pi}) + \sum_{s\in S_i}(p_i{}^*(a))_s r(s,a_s)$$
$$+\beta\{\sum_{j=1}^{n}[\sum_{s'\in S_j}\sum_{s\in S_i}(p_i{}^*(a))_s d(s'/s,a_s)]\}\overline{V}_\mu(j,\overline{\pi})$$
$$= \beta\overline{V}_\mu(i,\overline{\pi}) + \sum_{s\in S_i}(p_i{}^*(a))\{r(s,a_s) + \beta\sum_{j=1}^{n}\sum_{s'\in S_j} d(s'/s,a_s)\overline{V}_\mu(j,\overline{\pi})\}\ .$$

We can consider that:

$r(s,a_s) + \beta \sum_{j=1}^{n}\sum_{s'\in S_j} d(s'/s,a_s)\overline{V}_\mu(j,\overline{\pi})$ is some reward in Γ_i , which results from the choice of action $\boldsymbol{a_s}$ if the process is in state s.

If we set : $\overline{c}_i(s,a_s) = r(s,a_s) + \beta \sum_{j=1}^{n}\sum_{s'\in S_j} d(s'/s,a_s)\overline{V}_\mu(j,\overline{\pi})$; it follows that:

$$c(i,a) + \beta\sum_{j=1}^{n} q(j/i,a)\overline{V}_\mu(j,\overline{\pi}) = \beta\overline{V}_\mu(i,\overline{\pi}) + (p_i{}^*(a))\overline{C}_i(a)^T, \qquad (4.2)$$

where $\overline{C}_i(a) = (\overline{c}_i(s,a_s))_{s\in S_i}$.

Note that: $(p_i{}^*(a))\overline{C}_i(a)^T$ is the value of the strategy $\boldsymbol{a : (s \to a_s, s \in S_i)}$,

in the irreducible MDP Γ_i in which the rewards are defined by ; $\overline{c}_i(s,a_s)$.

From (4.1) and (4.2), it follows that the problem defined in step2 of Algorithm1 is similar to:

$$\beta\overline{V}_\mu(i,\overline{\pi}) + (p_i{}^*(a))\overline{C}_i(a)^T > \beta\overline{V}_\mu(i,\overline{\pi}) + p_i{}^*(\overline{\pi}(i))\overline{C}_i(\overline{\pi}(i))^T\ .$$

Hence, the problem is to find an action $\boldsymbol{a}$ such that:

$$p_i{}^*(a)\overline{C}_i(a)^T > p_i{}^*(\overline{\pi}(i))\overline{C}_i(\overline{\pi}(i))^T\ . \qquad (4.3)$$

It can be seen that the problem in (4.3) can be solved by one iteration of policy improvement algorithm (the initial strategy is $\overline{\pi}(i)$) to irreducible MDP Γ_i where the rewards are defined by $\overline{c}_i$.

We can now state the following algorithm for searching an action $a \in \overline{A}(i), i \in \overline{S}$, of step2 in Algorithm 1.

Algorithm 3

a) Fix $i \in \overline{S}$ and $\overline{\pi}(i)$.

b) Compute $\lambda, y \in \mathbf{R}^{n_i}$, solution of the linear system:

$\lambda e^T + y = \overline{C}_i(\overline{\pi}(i) + P_i(\overline{\pi}(i)y$ with $y_{n_i} = 0$ where $y = (y_1, y_2, ..., y_{n_i})^T \in \mathbf{R}^{n_i}$ and $(1, 1, ..., 1) \in \mathbf{R}^{n_i}$ and $P_i(\overline{\pi}(i) = [p_{ss'}(\overline{\pi}(i)]_{s,s' \in S_i}$.

c) Find for each s in S_i, an action a_s in $A(s)$ such that:

$$\left(\overline{c}_i(s, a_s) + \sum_{s' \in S_i} p(s'/s, a_s) y_{s'}\right) > (\lambda + y_s)\,.$$

d) If $\boldsymbol{a_s}$ does not exist, for every s de S_i, then (stop), otherwise go to e)

e) Choose $\overline{\pi}^1(i)$ deterministic such that:

$\overline{\pi}^1(i)(s) = a_s$ if a_s exists and $\overline{\pi}^1(i)(s) = \overline{\pi}(i)(s)$ if a_s does not exist; $(s \in S)$

h) Let $a := (\overline{\pi}^1(i))(s)$

From the previous results, our aggregation-disaggregation algorithm for solving the limit M.C problem (L) is stated as follows:

Step1 Select an arbitrary deterministic strategy $\pi \in \Gamma$, the corresponding strategy $\overline{\pi} \in \overline{\Gamma}$ is defined by:

$\overline{\pi}(i) = (\pi(s))_{s \in S_i}, i \in \overline{S}$

Step2 Apply Algorithm2 to compute $[p_i^*(\overline{\pi}(i))], i \in \overline{S}$.

Step3 Compute

$$C(\overline{\pi}) = (c(i, \overline{\pi}(i))_{i \in \overline{S}} \text{ and } Q(\overline{\pi}) = (q(j/i, \overline{\pi}(i)))_{i,j \in \overline{S}}\ .$$

Step4 Compute: $\overline{V}_\mu(\overline{\pi}) = [I_n - \frac{1}{1+\mu} Q(\overline{\pi})]^{-1} C(\overline{\pi})$.

Step5 For each $i \in \overline{S}$ and for $\overline{\pi}(i)$, apply Algorithm3 for searching an action $a \in \overline{A}(i)$.

Step 6 Define the strategy : for all $i \in \overline{S}$, $\overline{\pi}_1(i) = \begin{cases} a & if\ a\ exists \\ \\ \overline{\pi}(i) & if\ a\ does\ not\ exist\ . \end{cases}$

Step7 If $\overline{\pi}_1 = \overline{\pi}$, then $\overline{\pi}$ is optimal (stop), otherwise set $\overline{\pi} := \overline{\pi}_1$, and go to step 2.

References

[1] M.Abbad, Jerzy A .Filar (1992) , “Perturbation and Stability theory for Markov Control Problems”. *IEEE Tranc. Automat. Control*, Vol.37, No . 9.

[2] M.Abbad and Jerzy A Filar (1995), ”Algorithms for Singularly Perturbed Markov Control Problems : A Survey”. *Control and Dynamic Systems*, Vol.73, Academic Press.

[3] M.Abbad and Jerzy A Filar (1991), “Perturbation Theory for Semi Markov Control Problems”. In: *Proceedings of the 30th Conference on Decision and Conrol*, December 1991 (England).

[4] M.Abbad and M.Kissai (1995), “ linear Programming Algorithm for Singularly Perturbed Discounted Markov Control Problems”, Preprint, 1995.

[5] M.Abbad , Jerzy A .Filar , and Tomasz R. Bielecki (1992), “Algorithms For singularly Perturbed Limiting Average Markov Control Problems”. *IEEE Tranc. Automat. Control*, Vol.37,No . 9.

[6] R.Aldhaheri and H.Khalil (1989), “Aggregation and optimal control of nearly completely decomposable markov chains”. In: *Proceedings of 28th CDC*, IEEE, pp.1277-1282.

[7] Arthur F.Veinott . JR. (Stanford University) (1969), “ Discrete Dynamic Programming with sensitive Discount Optimality Criteria ”.

[8] T.R.Bielecki and J.A . Filar (1991), “ Singularly Perturbed Markov Control Problem: Limiting Average Cost ”. *Annals of Operations Research*, Vol.28, pp.153-168.

[9] D.Blackwell (1962), ”Discrete Dynamic Programing”. *Annals of Mathematical Statistics*, 33, pp.719-726

[10] P.J Courtois (1977), *Decomposability :Queueing and Computing Systems*, academic Press, N.Y .

[11] F.Delebecque (1983), “ A Reduction Process for Perturbed Markov Chains”. *SIAM J. App. Math.*, vol AC - 48 , pp.325 -350.

[12] F.Delebecque and J.Quadrat (1981), “ Optimal control of Markov chains admitting strong and weak interractions ”. *Automatica*, 17 , pp.281-296.

[13] E.V Denardo (1982), *Dynamic Programming*, Prentice- Hall Egle- Wood Cliffs , New Jersey .

[14] C.Derman (1970), *Finite State Markovian Decision Process*, Academic Press, N.Y.

[15] L.C.M. Kallenberg (1983), “ Linear Programming and Finitte Markovian Control Problem ”. *Mathematical Center Tracts*, 148 , Amsterdam .

[16] T.Kato (1980), *Perturbation Theory for Linear Operators*, Spring-Verlag , Berlin .

[17] J. Kemeny and J.L.Snell (1960), *Finite Markov Chains*, Van Nostrand , New York .

[18] Paul J.Schweitzer (1986), "Perturbation series expansions for nearly completely-decomposable Markov Chains". *Teletrafic Analysis and computer performance evaluation*, pp. 319-328.

[19] R.G Phillips and P.Kokotovic (1981), " A Singular Perturbation Approch to Modeling and Control of Markov Chains". *IEEE Transactions on Automatic Control*, (AC-26) (pp.1087-1094).

[20] Pravin, Varaiya (1978), "Optimal and suboptimal stationary controls for Markov Chains". *IEEE Transactions on Automatic control*, vol AC 23 No 3 , June 1978.

[21] Ronald A. Howard (1960), *Dynamic Programming and Markov Processes.* The M.I.T. Press Massachusetts institute of Tecnology Cambridge , Massachusetts N.Y.

[22] Sheldon M. Ross (1971), *Applied Probability Models with Optimization Applications.* (Holden-day.San Francisco).

[23] Sheldon M. Ross (1983), *Introduction to Stochastic Dynamic Programming.* N.Y: Academic 1983.

[24] H.A. Simon and A.Ando (1961), " Aggregation of variables in Dynamicc Systems ". *Econometrica*, vol .29, 111-138.