Proceedings of the 29th Conference
on Decision and Control
Honolulu, Hawaii • December 1990

Algorithms for Singularly Perturbed Limiting Average Markov Control Problems [1]

Mohammed Abbad and Jerzy A. Filar
Department of Mathematics and Statistics
University of Maryland at Baltimore County
Baltimore, Maryland

Tomasz R. Bielecki
Department of Mathematics
University of Kansas
Lawrence, Kansas

1. Introduction

In this paper we consider a singularly perturbed Markov Decision Process with the limiting average cost criterion. We assume that the underlying process is composed of n separate irreducible processes, and that the small perturbation is such that it "unites" these processes into a single irreducible process. This structure corresponds to the Markov chains admitting "strong and weak interactions" that arises in many applications, and was studied by a number of authors (e.g., see Delebecque and Quadrat [5], Phillips and Kokotovic [13], Coderch et al. [3], Kokotovic [10], Schweitzer [15] and [16], Rohlicek and Willsky [14], and Aldhaheri and Khalil [1]). Our results can also be viewed as a continuation of a line of research initiated by Schweitzer [15] in 1968.

In Section 2 we introduce the formulation and some results given by Bielecki and Filar [2] of the underlying control problem for the singularly perturbed MDP; the so-called "limit Markov Control Problem" (limit MCP). In particular these authors proved that an optimal solution to the perturbed MDP can be approximated by an optimal solution of the limit MCP for sufficiently small perturbation.

In Section 3 we demonstrate that the above limit Markov Control Problem can be solved by a suitably constructed linear program.

In Section 4 we construct an algorithm for solving the limit Markov Control Problem based on the policy improvement method. Recently we learned that this algorithm is similar to one given by Pervozvanskii and Gaitsgori [12]. However, these authors did not explicitly consider the limit Markov Control Problem, and worked only in the smaller class of deterministic strategies.

2. Definitions and Preliminaries

A discrete Markovian Decision Process (MDP, for short) is observed at time points $t = 0, 1, 2, \ldots$. The *state space* is denoted by $S = \{1, 2, \ldots, N\}$. With each state $s \in S$ we associate a finite *action set* $A(s) = \{1, 2, \ldots, m_s\}$.[1] At any time point t the system 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 a state $s' \in S$ with transition probability $p(s' \mid s,a)$, where $p(s' \mid s,a) \geq 0$ and $\sum_{s' \in S} p(s' \mid s,a) = 1$. Henceforth, such an MDP will be synonymous with the four-tuple:
$\Gamma = \langle S, \{A(s); s \in S\}, \{r(s,a); (s,a) \in S \times A(s)\}, \{p(s' \mid s,a); (s,a,s') \in S \times A(s) \times S\}\rangle$.
While a general control strategy in Γ may depend on the complete state-action histories of the process, in this paper we shall concern ourselves only with the class Π of all stationary strategies. A *stationary strategy* $\pi \in \Pi$ is the vector:

$$\pi = \{\pi(s,a) \mid (s,a) \in S \times A(s)\},$$

where $\pi(s,a)$ is the probability that the 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 $\pi \in \Pi$ will be called *deterministic* if $\pi(s,a) \in \{0,1\}$ for all $(s,a) \in S \times A(s)$.

With every $\pi \in \Pi$ we shall associate the following quantities: $\mathbf{r}(\pi) = \big(r_1(\pi), \ldots, r_N(\pi)\big)^T$, 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 $P(\pi) = \big(p_{ss'}(\pi)\big)_{s,s'=1}^{N}$, where $p_{ss'}(\pi) := \sum_{a \in A(s)} p(s' \mid s,a)\pi(s,a)$ 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* (sometimes called the *ergodic projection at infinity*) defined by:

$$P^*(\pi) = \big(p^*_{ss'}(\pi)\big)_{s,s'=1}^{N} := \lim_{t \to \infty} \frac{1}{t+1} \sum_{k=0}^{t} P^k(\pi),$$

where $P^0(\pi) := I_N$, an $N \times N$ identity matrix; the overall performance index resulting from the use of π, namely, $\mathcal{J}(s,\pi) := \big[P^*(\pi)\mathbf{r}(\pi)\big]_s$ for each initial state $s \in S$.[2]

The "classical" limiting average Markov Decision problem is the optimization problem: Find $\pi^0 \in \Pi$ such that

$$\mathcal{J}(s,\pi^0) = \max_{\pi} \mathcal{J}(s,\pi) \qquad \text{for all } s \in S.$$

A strategy π^0 satisfying the above will be called *optimal*. It is well known that there always exists an optimal deterministic policy and there is a number of finite algorithms for its computation (e.g., Denardo [6], Derman [7], Kallenberg [8]).

In this paper we shall assume that:

(A1) $S = \cup_{i=1}^{n} S_i$ where $S_i \cap S_j = \emptyset$ if $i \neq j$, $n > 1$, $\text{card} S_i = n_i$, $n_1 + \ldots + n_n = N$,

and

(A2) $p(s' \mid s,a) = 0$ whenever $s \in S_i$ and $s' \in S_j$, $i \neq j$.

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 $i = 1, 2, \ldots, n$, respectively. Note that if Π_i is the space of stationary strategies in Γ_i, then a strategy $\pi \in \Pi$ in Γ can be written in the natural way as $\pi = (\pi^1, \pi^2, \ldots, \pi^n)$, where $\pi^i \in \Pi_i$. The probability transition matrix in Γ_i corresponding to π^i is, of course, defined by: $P_i(\pi^i) := \big(p_{s,s'}(\pi^i)\big)_{s,s' \in S_i}$, and the generator $G_i(\pi^i)$ and the Cesaro-limit $P_i^*(\pi^i)$ matrices can be defined in a manner analogous to that in the original process Γ. In addition, we assume that:

[1] Note that action $i \in A(s)$ may not be the same as action $i \in A(s')$ if $s \neq s'$. This simplification of notation should not cause ambiguity.

[2] The notation $[\mathbf{u}]_s$ will be used to denote the s-th entry of a vector $\mathbf{u}$.

[1] We are indebted to Marc Teboulle for pointing out reference [10] to us.

(A3) For every $i = 1, 2, \ldots, n$ and for all $\pi^i \in \Pi_i$ the matrix $P_i(\pi^i)$ is an irreducible matrix.

In view of (A3), $P_i^*(\pi^i)$ is a matrix with identical rows. We shall denote any row of $P_i^*(\pi^i)$ by $\mathbf{p}_i^*(\pi^i)$.

Remark 2.1 *Note that for all $\pi \in \Pi$ we have the following representation of $P^*(\pi)$:*

$$P^*(\pi) = EM(\pi)$$

where E is an $N \times n$ matrix with entries:

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

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

$$m_{js}(\pi) = \begin{cases} [\mathbf{p}_j^*(\pi^j)]_s & \text{if } \sum_{k=1}^{j-1} n_k < s \leq \sum_{k=1}^{j} n_k \\ 0 & \text{otherwise} \end{cases}$$

for $j = 1, 2, \ldots, n$ and $s = 1, 2, \ldots, N$. Of course we set $\sum_{k=1}^{0} n_k := 0$. Note also that from the above definitions we conclude that:

$$M(\pi)E = I_{n \times n}.$$

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:

$$D = \{d(s' \mid s, a) \mid (s, a, s') \in S \times A(s) \times S\}$$

where the elements of D satisfy $\sum_{s' \in S} d(s' \mid s, a) = 0$ for all $(s, a) \in S \times A(s)$,
$-1 \leq d(s \mid s, a) \leq 0$, $d(s' \mid s, a) \geq 0$, and $s \neq s'$.

Now, with every $\pi \in \Pi$ we can associate a *perturbation generator matrix* $D(\pi) = \left(d_{ss'}(\pi)\right)_{s,s'=1}^{N}$ where $d_{ss'}(\pi) = \sum_{a \in A(s)} d(s' \mid s, a)\pi(s, a)$. We shall also require that there exists $\epsilon_0 > 0$ such that for every $\pi \in \Pi$,

$$G_\epsilon(\pi) := G(\pi) + \epsilon D(\pi) \tag{2.1}$$

is a generator of a Markov chain for any $\epsilon \leq \epsilon_0$.

We shall consider a family of *perturbed processes* Γ_ϵ for $\epsilon \in [0, \epsilon_0]$ that differ from the original MDP Γ only in the transition law, namely, in Γ_ϵ for every $(s, a, s') \in S \times A(s) \times S$ we have:

$$p_\epsilon(s' \mid s, a) := p(s' \mid s, a) + \epsilon d(s' \mid s, a). \tag{2.2}$$

By (2.1) and (2.2) we have that every $\pi \in \Pi$ induces in the perturbed process Γ_ϵ the Markov Chain with the probability transition matrix:

$$P_\epsilon(\pi) = G_\epsilon(\pi) + I_N. \tag{2.3}$$

The most important structural assumption of this paper is stated below.

(SPA) Singular Perturbation Assumption:

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

Remark 2.2 *Note that as a result of (SPA) and (A1)-(A3) the rank of $P_\epsilon^*(\pi)$ is 1, which is strictly less than n, the rank of $P^*(\pi)$. Consequently our perturbation is indeed a singular perturbation in the sense of Delebecque [4].*

Under assumption SPA the MDP Γ_ϵ, $\epsilon \in [0, \epsilon_0]$, defined as:
$\Gamma_\epsilon = \langle S, \{A(s); s \in S\}, \{r(s, a); (s, a) \in S \times A(s)\}, \{p_\epsilon(s' \mid s, a); (s, a, s') \in S \times A(s) \times S\} \rangle$
is called the *singularly perturbed MDP.*

Denoting by $P_\epsilon^*(\pi)$ the ergodic projection at infinity corresponding to $P_\epsilon(\pi)$ we define the overall performance index resulting from the use of $\pi \in \Pi$ by:

$$\mathcal{J}_\epsilon(s, \pi) = [P_\epsilon^*(\pi)\mathbf{r}(\pi)]_s, \quad s \in S.$$

The optimal value function $\mathcal{J}_\epsilon$ corresponding to Γ_ϵ is given by:

$$\mathcal{J}_\epsilon(s) = max_{\pi \in \Pi} \mathcal{J}_\epsilon(s, \pi), \quad s \in S.$$

Note that by SPA, $P_\epsilon(\pi)$ is irreducible and hence there is no loss of generality in omitting the argument s in the above, and writing simply $\mathcal{J}_\epsilon(\pi)$, and $\mathcal{J}_\epsilon$ respectively.

We now recall from [2] the analysis of the limiting behavior of $\mathcal{J}_\epsilon$ as ϵ goes to zero and the validity of the so called *limit control principle* in the present framework, that is, an optimal strategy in a certain limit optimization problem, to be defined in sequel, is δ-optimal in Γ_ϵ for any $\delta > 0$ and for ϵ sufficiently small.

The problems described above pertain to the class of MDP's with singularly perturbed transition kernels or more generally to the class of control problems for Markov Processes with singularly perturbed generators. Related problems for discounted cost criteria have been addressed in Delebecque & Quadrat [5] and Delebecque [4] in the case of discrete time Markov chains and in Philips & Kokotovič [13] for both continuous and discrete time Markov chains.

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

$$\bar{B}(\pi) \stackrel{\text{def}}{=} M(\pi)D(\pi)E.$$

We note that by Remark 2.1 $\bar{B}(\pi)$ is a generator of an "aggregated" Markov chain on a state space $\bar{S} \stackrel{\text{def}}{=} \{1, 2, \ldots, n\}$.

Remark 2.3 *By A1-A3 and SPA the operator $\bar{B}(\pi)$ defines an irreducible Markov chain on $\bar{S}$. This can be verified by direct inspection, (see also remark 2 p. 338 in Delebecque [4]).*

Now let $\bar{P}^*(\pi)$ denote the ergodic projection at infinity corresponding to $\bar{B}(\pi)$, for each $\pi \in \Pi$. Define, for $\pi \in \Pi$, an $N \times N$ matrix:

$$\hat{P}^*(\pi) = E\bar{P}^*(\pi)M(\pi). \tag{2.4}$$

From Delebecque [4] Theorem 3, and the results of Kato [9], chapter II, Bielecki and Filar [2] obtained an asymptotic result stated in Lemma 2.1.

Lemma 2.1 *Under the assumptions A1-A3 and SPA*

$$\lim_{\epsilon \to 0}[\max_{\pi \in \Pi} \|P_\epsilon^*(\pi) - \hat{P}^*(\pi)\|] = 0, \tag{2.5}$$

where $\|.\|$ denotes any matrix norm.

Using this Lemma, the following Corollary was derived in [2].

Corollary 2.1 *Assume A1–A3 and SPA. Then for all $s \in S_j$, $j = 1, 2, \ldots, n$.*

$$\lim_{\epsilon \to 0} | \max_{\pi \in \Pi} [\bar{P}^*(\pi)\bar{\mathbf{r}}(\pi)]_j - \max_{\pi \in \Pi} [P^*_\epsilon(\pi)\mathbf{r}(\pi)]_s | = 0 \qquad (2.6)$$

where $\bar{\mathbf{r}}(\pi) := M(\pi)\mathbf{r}(\pi)$.

Remark 2.4 *The optimization problem:*

$$\max_{\pi \in \Pi} [\hat{P}^*(\pi)\mathbf{r}(\pi)]_s, \quad s \in S \qquad (L)$$

is called the the limit Markov Control Problem.

The optimization problem:

$$\max_{\pi \in \Pi} [\bar{P}^*(\pi)\bar{\mathbf{r}}(\pi)]_j, \quad j = 1, 2, \ldots, n \qquad (AL)$$

is called the aggregated limit Markov Control Problem.

Remark 2.5 *Note that in view of (2.4) we have:*

$$[\hat{P}^*(\pi)\mathbf{r}(\pi)]_s = [\bar{P}^*(\pi)\bar{\mathbf{r}}(\pi)]_j \text{ for } s \in S_j,\ j = 1, 2, \ldots, n,\ \pi \in \Pi. \qquad (2.9)$$

It follows that any maximizing strategy π^0 for (AL) is also a maximizing strategy for (L) and vice-versa. The existence of a maximizing strategy π^0 is clear. In view of (2.7) and the irreducibility properties it is also clear that π^0 does not depend on the initial state $s \in S$.

Using the results above, the following theorem was proved in [2]:

Theorem 2.1 *(Limit Control Principle) Assume A1–A3 and SPA. Let $\pi^0 \in \Pi$ be any maximizer in (AL). Then for all $\delta > 0$, there exists $\epsilon_\delta > 0$ such that for all $\epsilon < \epsilon_\delta$.*

$$\| P^*_\epsilon(\pi^0)\mathbf{r}(\pi^0) - \mathcal{J}_\epsilon \| \leq \delta. \qquad (2.8)$$

3. Linear Programming and the Limit Markov Control Problem

We have seen in Section 2 that the optimization problem:

$$\max_{\pi \in \Pi} [\hat{P}^*(\pi)\mathbf{r}(\pi)]_s, \quad s \in S \qquad (L)$$

is the natural problem to attempt to solve in the case of a singularly perturbed Markov Decision Process. We shall demonstrate that this problem can be converted to an equivalent problem in the space of the long-run state-action frequencies. Towards this goal we shall now consider one of the irreducible MDP's:

$\Gamma_i = \langle S_i, \{A(s); s \in S_i\}, \{r(s,a); (s,a) \in S_i \times A(s)\}, \{p(s' \mid s,a); (s,a,s') \in S_i \times A(s) \times S_i\} \rangle$ that were already introduced in Section 2, with $i = 1, 2, \ldots, n$. The following results are well-known (e.g., see Kallenberg [8]). With each Γ_i we can associate a polyhedral set:

$$X_i = \{\mathbf{x}^i = \{x^i_{sa} | (s,a) \in S \times A(s)\} | \sum_{s \in S_i} \sum_{a \in A(s)} x^i_{sa} = 1;$$
$$\sum_{s \in S_i} \sum_{a \in A(s)} \left(\delta_{s,s'} - p(s' \mid s,a)\right) x^i_{sa} = 0,\ s' \in S_i;$$
$$x^i_{sa} \geq 0,\ (s,a) \in S_i \times A(s)\};$$

and a bijective map $T : X_i \to \Pi_i$ such that $T(\mathbf{x}^i) := \pi^i$ where

$$\pi^i(s,a) := x^i_{sa} / \sum_{a \in A(s)} x^i_{sa};\ (s,a) \in S_i \times A(s).$$

Now, if $\pi^i \in \Pi_i$ and $\mathbf{p}^*_i(\pi^i)$ is the corresponding stationary distribution in Γ_i it can be shown that the inverse map is $T^{-1} : \Pi_i \to X_i$ such that $T^{-1}(\pi^i) := \mathbf{x}^i$, where

$$x^i_{sa} := [\mathbf{p}^*_i(\pi^i)]_s \pi^i(s,a);\ (s,a) \in S_i \times A(s).$$

We shall now rewrite the objective function of (L) (with the help of (2.4) and structural assumptions) as:

$$\hat{\mathcal{J}}(\pi) := [\hat{P}^*(\pi)\mathbf{r}(\pi)]_s = [E\bar{P}^*(\pi)M(\pi)\mathbf{r}(\pi)]_s$$
$$= \sum_{i=1}^n [\bar{\mathbf{p}}^*(\pi)]_i [\mathbf{p}^*_i(\pi^i) \cdot \mathbf{r}_i(\pi^i)], \qquad (3.1)$$

where $\bar{\mathbf{p}}^*(\pi)$ is the stationary distribution vector of $\bar{P}(\pi) = \bar{B}(\pi) + I_{n \times n}$, $s \in S$, and $\pi = (\pi^1, \pi^2 \ldots, \pi^n)$ as in Section 2.

It can be easily checked that with T as above we have:

$$[\mathbf{p}^*_i(\pi^i)]_s = [\mathbf{p}^*_i\left(T(\mathbf{x}^i)\right)]_s = \sum_{a \in A(s)} x^i_{sa} \qquad (3.2)$$

for each $i = 1, 2, \ldots, n$ and $s \in S_i$. Hence (3.1) - (3.2) yield:

$$\hat{\mathcal{J}}(\pi) = \sum_{i=1}^n [\bar{\mathbf{p}}^*(\pi)]_i \left(\sum_{s \in S_i} \sum_{a \in A(s)} r(s,a) x^i_{sa}\right) \qquad (3.3)$$

for every $s \in S$, where $\mathbf{x}^i = T^{-1}(\pi^i)$.

Recall that $\bar{\mathbf{p}}^*(\pi)$ is the unique solution of the system of equations:

$$\alpha^T \bar{P}(\pi) = \alpha^T$$
$$\sum_{i=1}^n \alpha_i = 1,$$

since $\bar{P}(\pi)$ is irreducible for all $\pi \in \Pi$. The first of these equations can, by definition of $\bar{B}(\pi)$, be expressed as:

$$\alpha^T M(\pi)(D(\pi) + I_N)E = \alpha^T \qquad (3.4)$$

Equivalently, using the fact that $M(\pi)I_N E = I_n$, we can transform (3.4) to:

$$\alpha^T M(\pi)D(\pi)E = \mathbf{0}^T. \qquad (3.5)$$

Note that $U(\pi) := M(\pi)D(\pi)E$ is an $n \times n$ matrix whose $(i,j)-$th entry is:

$$u_{ij}(\pi) := \sum_{s' \in S_j} \sum_{s \in S_i} [\mathbf{p}^*_i(\pi^i)]_s d_{ss'}(\pi^i). \qquad (3.6)$$

Now using (3.2) we see that with $\mathbf{x}^i := T^{-1}(\pi^i)$ for each $i = 1, 2, \ldots, n$, we have:

$$v_{ij}(\mathbf{x}) := u_{ij}(\pi) = \sum_{s' \in S_j} \sum_{s \in S_i} \sum_{a \in A(s)} d(s' \mid s,a) x^i_{sa}, \qquad (3.7)$$

where $\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n)$.

Setting $V(\mathbf{x}) = \left(v_{ij}(\mathbf{x})\right)_{i,j=1}^n$, we are now led to consider the following nonlinear programming problem (NL):

$$maximize \sum_{i=1}^n \sum_{s \in S_i} \sum_{a \in A(s)} r(s,a) x^i_{sa} \alpha_i$$

Subject to:

(i) $\mathbf{x}^i = \{x^i_{sa} \mid (s,a) \in S_i \times A(s)\} \in X_i,\ i = 1, 2, \ldots n,$

(ii) $\alpha_i \geq 0,\ i = 1, 2, \ldots, n,$ and $\sum_{i=1}^n \alpha_i = 1,$

(iii) $\alpha^T V(\mathbf{x}) = \mathbf{0}^T.$

Note that $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^n)$ and $\alpha = (\alpha_1, \ldots, \alpha_n)$ are the variables in the problem (NL).

We can now state the following result [2]:

Theorem 3.1 *Let $(\bar{\mathbf{x}}, \bar{\alpha})$ be an optimal solution of the nonlinear program (NL). Define $\bar{\pi}^i := T(\bar{\mathbf{x}}^i)$ for $i = 1, 2, \ldots, n$ and $\bar{\pi} := (\bar{\pi}^1, \ldots, \bar{\pi}^n)$, then $\bar{\pi}$ is optimal in the limit problem (L).*

Remark 3.6 *The importance of the above theorem stems from the fact that it converts the limit problem (L) into the problem (NL). In what follows, we shall demonstrate that an optimal solution in the program (NL) can be obtained from an optimal solution of an appropriate linear program that can be solved by efficient linear programming techniques.*

Consider the following linear programming problem (P):

$$\text{maximize } \sum_{i=1}^{n} \sum_{s \in S_i} \sum_{a \in A(s)} \mathbf{r}(s,a) z^i_{sa}$$

Subject to:

$$\sum_{s \in S_i} \sum_{a \in A(s)} \Big(\delta_{ss'} - p(s' \mid s,a)\Big) z^i_{sa} = 0; \quad s' \in S_i; \quad i = 1, \ldots, n \quad (3.8)$$

$$\sum_{i=1}^{n} \sum_{s' \in S_j} \sum_{s \in S_i} \sum_{a \in A(s)} d(s' \mid s,a) z^i_{sa} = 0; \quad j = 1, 2, \ldots, n \quad (3.9)$$

$$\sum_{i=1}^{n} \sum_{s \in S_i} \sum_{a \in A(s)} z^i_{sa} = 1 \quad (3.10)$$

$$z^i_{sa} \geq 0; \quad i = 1, 2, \ldots, n; \quad s \in S_i; \quad a \in A(s) \quad (3.11)$$

Lemma 3.1 *For any feasible solution $\mathbf{z}$ of (P),*

$$\sum_{s \in S_i} \sum_{a \in A(s)} z^i_{sa} > 0 \quad \textit{for all} \quad i \in \{1, 2, \ldots, n\}$$

Proof: Define $F(\mathbf{z}) := \{i \in \{1, 2, \ldots, n\} : \sum_{s \in S_i} \sum_{a \in A(s)} z^i_{sa} > 0\}$ and $\bar{F}(\mathbf{z}) := \{i \in \{1, 2, \ldots, n\} : \sum_{s \in S_i} \sum_{a \in A(s)} z^i_{sa} = 0\}$.
We shall show that $\bar{F}(\mathbf{z}) = \phi$. Assume $\bar{F}(\mathbf{z}) \neq \phi$. For $i \in \bar{F}(\mathbf{z})$ set $\alpha_i := 0$ and take any strategy $\pi^i \in \Pi_i$ and define $\mathbf{x}^i := T^{-1}(\pi^i)$. For $i \in F(\mathbf{z})$ define:

$$\alpha_i := \sum_{s \in S_i} \sum_{a \in A(s)} z^i_{sa} \quad (3.12)$$

and

$$x^i_{sa} := \frac{z^i_{sa}}{\sum_{s \in S_i} \sum_{a \in A(s)} z^i_{sa}} \quad \textit{for all} \quad s \in S_i, \; a \in A(s). \quad (3.13)$$

Note that we have:

$$z^i_{sa} = x^i_{sa} \alpha_i \quad \textit{for all} \quad i \in \{1, 2, \ldots, n\}, \quad s \in S_i, \quad a \in A(s). \quad (3.14)$$

From (3.8) and (3.13) we have: for all $i \in F(\mathbf{z})$

$$\sum_{s \in S_i} \sum_{a \in A(s)} \Big(\delta_{ss'} - p(s' \mid s,a)\Big) x^i_{sa} = 0 \quad \textit{for all} \quad s' \in S_i.$$

From (3.13) it follows that:

$$x^i_{sa} \geq 0 \quad \textit{for all} \quad s \in S_i, \quad a \in A(s), \quad \textit{and} \quad \sum_{s \in S_i} \sum_{a \in A(s)} x^i_{sa} = 1$$

It follows that for all $i \in F(z)$, $\mathbf{x}^i \in X_i$.
Since T is bijective, there exists a strategy π^i in Γ_i such that $\mathbf{x}^i = T^{-1}(\pi^i)$.
Now (3.9) and (3.14) imply that for all $j = 1, 2, \ldots, n$:

$$0 = \sum_{i=1}^{n} \sum_{s' \in S_j} \sum_{s \in S_i} \sum_{a \in A(s)} d(s' \mid s,a) x^i_{sa} \alpha_i$$

$$= \sum_{i=1}^{n} \alpha_i [\sum_{s' \in S_j} \sum_{s \in S_i} \sum_{a \in A(s)} d(s' \mid s,a) [p^*_i(\pi_i)]_s \pi^i(s,a)]$$

$$= \sum_{i=1}^{n} \alpha_i [\sum_{s' \in S_j} \sum_{s \in S_i} [p^*_i(\pi^i)]_s d_{ss'}(\pi^i)] = [\alpha^T M(\pi) D(\pi) E]_j,$$

where $\alpha^T = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and $\pi = (\pi^1, \pi^2, \ldots, \pi^n)$.

Thus $\alpha^T M(\pi)(D(\pi) + I_N)E = \alpha^T$ since $M(\pi) I_N E = I_n$, which is the same as $\alpha^T \bar{P}(\pi) = \alpha^T$.
From the definitions of α_i we have by (3.10):

$$\sum_{i=1}^{n} \alpha_i = \sum_{i=1}^{n} \sum_{s \in S_i} \sum_{a \in A(s)} z^i_{sa} = 1,$$

and by (3.11) $\alpha_i \geq 0$ for $i = 1, \ldots, n$.
Now we have:

$$\begin{aligned} \alpha^T \bar{P}(\pi) &= \alpha^T \\ \sum_{i=1}^{n} \alpha_i &= 1; \quad \alpha_i \geq 0; \quad i = 1, \ldots, n. \end{aligned}$$

But $\bar{P}(\pi)$ is irreducible, hence all α_i must be positive, and thus we have a contradiction. Therefore $\bar{F}(z)$ must be empty. □

Theorem 3.2 *If $\bar{z}$ is an optimal solution of (P), then $(\bar{\mathbf{x}}, \bar{\alpha})$ is an optimal solution of (NL). Where*

$$\bar{x}^i_{sa} := \frac{\bar{z}^i_{sa}}{\sum_{s \in S_i} \sum_{a \in A(s)} \bar{z}^i_{sa}} \quad \textit{and} \quad \bar{\alpha}_i := \sum_{s \in S_i} \sum_{a \in A(s)} \bar{z}^i_{sa}$$

for $i = 1, 2, \ldots, n$; $s \in S_i$; $a \in A(s)$.

Proof: By Lemma 3.1, $\bar{\mathbf{x}}$ is well defined. Also, from the proof of Lemma 3.1, it follows that $(\bar{\mathbf{x}}, \bar{\alpha})$ is a feasible solution in (NL). To prove that $(\bar{\mathbf{x}}, \bar{\alpha})$ is optimal in (NL), let $(\mathbf{x}, \alpha)$ be any feasible solution in (NL) and define $\mathbf{z}$ by $z^i_{sa} := x^i_{sa} \alpha_i; \quad i = 1, \ldots, n; \quad s \in S_i; \quad a \in A(s)$. Note that $\mathbf{z}$ is feasible for (P), and we have:

$$\sum_{i=1}^{n} \sum_{s \in S_i} \sum_{a \in A(s)} r(s,a) x^i_{sa} \alpha_i = \sum_{i=1}^{n} \sum_{s \in S_i} \sum_{a \in A(s)} r(s,a) z^i_{sa} \leq$$

$$\sum_{i=1}^{n} \sum_{s \in S_i} \sum_{a \in A(s)} r(s,a) \bar{z}^i_{sa} = \sum_{i=1}^{n} \sum_{s \in S_i} \sum_{a \in A(s)} \bar{x}^i_{sa} \bar{\alpha}_i.$$

This proves the optimality of $(\bar{\mathbf{x}}, \bar{\alpha})$. □

Next we shall show that an optimal deterministic strategy for the limit Markov Control Problem (L) can be constructed from an extreme optimal solution of the linear programming problem (P).

Lemma 3.2 *Let $\bar{\mathbf{z}}$ be an extreme feasible solution for (P), then for any $i \in \{1, 2, \ldots, n\}$ and any $s \in S_i$ there is a unique $a \in A(s)$ such that $\bar{z}^i_{sa} > 0$.*

Proof: Since in particular $\bar{\mathbf{z}}$ is a feasible solution for (P), it follows from the proof of Lemma 3.1 that for any $i \in \{1, 2, \ldots, n\}$ there exists a strategy π^i such that:

$\frac{\bar{z}^i_{sa}}{\sum_{s\in S_i}\sum_{a\in A(s)}\bar{z}^i_{sa}} = [T^{-1}(\pi^i)]_{sa} := [\mathbf{p}^*_i(\pi^i)]_s \pi^i(s,a), \quad s \in S_i,$
$a \in A(s)$.

Hence for all $i \in \{1,2,\ldots,n\}$, $\quad s \in S_i$, there exists $a \in A(s)$: $\bar{z}^i_{sa} > 0$. Thus the number of positive elements in $\bar{\mathbf{z}}$ is at least $n_1 + n_2 + \ldots + n_n := N$. On the other hand, since $\bar{\mathbf{z}}$ is an extreme feasible solution for (P), the number of positive elements in $\bar{\mathbf{z}}$ is not greater than the rank k of the matrix corresponding to the linear program (P). From that matrix we have: for any $i \in \{1,2,\ldots,n\}$, $\sum_{s'\in S_i}|\sum_{s\in S_i}\sum_{a\in A(s)}\big(\delta_{ss'} - p(s' \mid s,a)\big)z^i_{sa}| = \sum_{s\in S_i}\sum_{a\in A(s)}(1-1)z^i_{sa} = 0$, and $\sum_{j=1}^n[\sum_{i=1}^n\sum_{s'\in S_j}\sum_{s\in S_i}\sum_{a\in A(s)} d(s' \mid s,a)z^i_{sa}]$
$= \sum_{i=1}^n\sum_{s\in S_i}\sum_{a\in A(s)}[\sum_{j=1}^n\sum_{s'\in S_j} d(s' \mid s,a)]z^i_{sa} = 0$.

Hence $k \le (n_1 - 1) + (n_2 - 1) + \ldots + (n_n - 1) + (n-1) + 1 = n_1 + n_2 + \ldots + n_n := N$. Therefore the number of positive elements in $\bar{z}$ is exactly N, and we conclude that $\bar{z}^i_{sa} > 0$ for exactly one $a \in A(s)$ for any $i \in \{1,2,\ldots,n\}$ and any $s \in S_i$. □

Theorem 3.3 *The limit Markov Control Problem (L) has an optimal deterministic strategy.*

Proof: It is clear that the problem (NL) is feasible. This implies that the problem (P) is feasible (if $(\mathbf{x},\alpha)$ is feasible for (NL), then $z^i_{sa} = x^i_{sa}\alpha_i$ is feasible for (P)). Since the constraints in (P) define a bounded polyhedron, then (P) has an optimal extreme solution, say $\bar{\mathbf{z}}$. From Theorem 3.2 the corresponding $(\bar{\mathbf{x}},\bar{\alpha})$ is an optimal solution for (NL). From Theorem 3.1, $\bar{\pi} := (\bar{\pi}^1,\ldots,\bar{\pi}^n)$ where $\bar{\pi}^i := T(\bar{\mathbf{x}}^i)$ for $i = 1,2,\ldots,n$ is optimal in (L). From Lemma 3.2, π must be deterministic. □

Remark 3.7 *Note that from Lemma 3.2 and from the proof of Theorem 3.3, it follows that if $\mathbf{z} := \{z^i_{sa} \mid s \in S_i; i = 1,\ldots,n; a \in A(s)\}$ is an optimal extreme solution of the linear program (P), then the policy defined by:*

$$f(s) = a; \quad s \in S_i; \quad i = 1,\ldots,n \iff z^i_{sa} > 0$$

is optimal in the limit Markov control problem (L).

Remark 3.8 *In this section we proved that the limit problem (L) can be converted into the linear program (P). In view of the special structure of (P), an approach of the Wolfe-Dantzig decomposition method (e.g., see Murty [11]) is suggested. Hence the linear program (P) may not suffer from the curse of dimensionality. In the next section we shall construct another algorithm which does not suffer from the dimensionality of the problem.*

4. Aggregation-Disaggregation Algorithm.

Consider the Aggregated Markov Decision Process $\bar{\Gamma}$ defined as follows:

The State Space of $\bar{\Gamma}$: $\bar{S} := \{1,2,\ldots,n\}$
The Action Spaces of $\bar{\Gamma}$: $\bar{A}(i) := X_{s\in S_i}A(s)$ *for each* $i \in \bar{S}$
The Transition Law of $\bar{\Gamma}$: *for all* $(i,\mathbf{a},j) \in \bar{S}\times\bar{A}(i)\times\bar{S}$,

$$q_{ij}(\mathbf{a}) := \begin{cases} 1 + \sum_{s'\in S_i}\sum_{s\in S_i}[\mathbf{p}^*_i(\mathbf{a})]_s d(s' \mid s,a_s) & i = j \\ \sum_{s'\in S_j}\sum_{s\in S_i}[\mathbf{p}^*_i(\mathbf{a})]_s d(s' \mid s,a_s) & i \ne j \end{cases} \quad (4.1)$$

The Rewards of $\bar{\Gamma}$: *for all* $i \in \bar{S}$, $\mathbf{a} \in \bar{A}(i)$

$$c(i,\mathbf{a}) := \sum_{s\in S_i}[\mathbf{p}^*_i(\mathbf{a})]_s r(s,a_s), \quad (4.2)$$

where $\mathbf{a} = \{a_s \mid s \in S_i\}$.

Note that for any $i \in \bar{S}$, each action $\mathbf{a} \in \bar{A}(i)$ defines a deterministic strategy which maps $s \in S_i$ onto a_s. Thus in (4.1) and (4.2), $\mathbf{p}^*_i(\mathbf{a})$ is well defined.

The validity of the Transition Law, namely:

$\sum_{j=1}^n q_{ij}(\mathbf{a}) = 1, \quad i \in \bar{S}, \quad \mathbf{a} \in \bar{A}(i)$ and $q_{ij}(\mathbf{a}) \ge 0$,
$(i,\mathbf{a},j) \in \bar{S}\times\bar{A}(i)\times\bar{S}$
can be checked by inspection using the assumptions made on the disturbance law D.

The classical limiting average Markov Decision problem for $\bar{\Gamma}$ is the optimization problem

$$\max_{\bar{\pi}\in\bar{\Pi}}[Q^*(\bar{\pi})\mathbf{c}(\bar{\pi})]_i, \quad i \in \bar{S} \quad (\mathrm{AP})$$

where $\bar{\Pi}$ is the class of all stationary strategies in $\bar{\Gamma}$, $\quad Q^*(\bar{\pi})$ is the Cesaro-limit matrix corresponding to the Markov matrix $Q(\bar{\pi})$ which is defined by $Q(\bar{\pi}) := (q_{ij}(\bar{\pi}))_{i,j=1}^n$ where $q_{ij}(\bar{\pi}) := \sum_{\mathbf{a}\in\bar{A}(i)} q_{ij}(\mathbf{a})\bar{\pi}(i,\mathbf{a})$ for all $i,j \in \bar{S}$; and $\mathbf{c}(\bar{\pi}) := (c_1(\bar{\pi}),\ldots,c_n(\bar{\pi}))$ is the vector of single stage expected rewards in which $c_i(\bar{\pi}) := \sum_{\mathbf{a}\in\bar{A}(i)} c(i,\mathbf{a})\bar{\pi}(i,\mathbf{a})$ for each $i \in \bar{S}$.

Note that there is a one to one correspondence between the deterministic strategies in $\bar{\Gamma}$ and Γ.

Let $\bar{f}$ be a deterministic strategy in $\bar{\Gamma}$, the corresponding deterministic strategy f in Γ is defined by:

$$for \quad each\ i \in \bar{S}\ and\ s \in S_i, \quad f(s) := [\bar{f}(i)]_s.$$

Proposition 4.1 *Let $\bar{f}$ be an optimal deterministic strategy for the problem (AP), then the corresponding f is an optimal deterministic strategy for the limit Markov Control Problem (L)*

Proof: Note that from the theory of Markov Decision Processes (e.g., see [6]-[8]) $\bar{f}$ always exists. It can be verified by inspection that $Q(\bar{f}) = \bar{P}(f)$ since $\bar{P}(f) = I_n + M(f)D(f)E$. Thus $Q^*(\bar{f}) = \bar{P}^*(f)$ and $Q^*(\bar{f})\mathbf{c}(\bar{f}) = \bar{P}^*(f)\bar{\mathbf{r}}(f)$. Therefore f must be an optimal solution for the problem (AL), since the problem (AL) has an optimal deterministic strategy (Remark 2.5 and Theorem 3.3). Now from Remark 2.5 we conclude that f is an optimal strategy for the problem (L). □

Remark 4.9 *In view of the fact that for any deterministic strategy $\bar{f}$ in the process $\bar{\Gamma}$ we have $Q(\bar{f}) = \bar{P}(f)$ and $\bar{P}(f)$ is irreducible (Remark 2.3), thus $Q(\bar{f})$ is irreducible. This shows that $\bar{\Gamma}$ is an irreducible MDP and therefore it can be solved by using the simplified policy improvement algorithm (e.g., see Denardo [6], Derman [7], Kallenberg [8]).*

The "standard" policy improvement algorithm for the aggregated MDP $\bar{\Gamma}$ can be stated as follows:

Step 1: Select an arbitrary deterministic strategy $\bar{f}$.

Step 2: Solve the following linear system for the unknowns $\lambda, y_1, \ldots, y_{n-1}$:

$$\lambda + y_i = c\big(i,\bar{f}(i)\big) + \sum_{j=1}^n q_{ij}\big(\bar{f}(i)\big)y_j; \quad i = 1,2,\ldots,n \quad (4.3)$$

where $y_n := 0$.

Step 3: Find a deterministic strategy $\bar{g}$ that satisfies:

$$c\big(i,\bar{g}(i)\big)+\sum_{j=1}^{n} q_{ij}\big(\bar{g}(i)\big)y_j > c\big(i,\bar{f}(i)\big)+\sum_{j=1}^{n} q_{ij}\big(\bar{f}(i)\big)y_j i=1,2,\ldots,n. \tag{4.4}$$

If this is not possible for some $i \in \{1,2,\ldots,n\}$, then set $\bar{g}(i) = \bar{f}(i)$.

If $\bar{g} = \bar{f}$, STOP.

Otherwise $\bar{f} \leftarrow \bar{g}$ and go to Step 2.

Now, we shall show that for each $i = 1,2,\ldots,n$ the problem defined in (4.4) can be converted to one iteration of the policy improvement algorithm for an MDP defined by Γ_i except for the rewards which will be defined appropriately.

For every $i = 1,\ldots,n$ and $\mathbf{a} \in \bar{A}(i)$ we have:
$c(i,\mathbf{a}) + \sum_{j=1}^{n} q_{ij}(\mathbf{a})y_j = \sum_{s\in S_i}[\mathbf{p}_i^*(\mathbf{a})]_s r(s,a_s) + \sum_{\substack{j=1\\ j\neq i}}^{n}\{\sum_{s'\in S_j}$
$\sum_{s\in S_i}[\mathbf{p}_i^*(\mathbf{a})]_s d(s' \mid s,a_s)\}y_j + \{1+\sum_{s'\in S_i}\sum_{s\in S_i}[\mathbf{p}_i^*(\mathbf{a})]_s d(s' \mid s,a_s)\}y_i$

$= y_i + \sum_{s\in S_i}[\mathbf{p}_i^*(\mathbf{a})]_s r(s,a_s) + \sum_{j=1}^{n}\sum_{s'\in S_j}\sum_{s\in S_i}[\mathbf{p}_i^*(\mathbf{a})]_s d(s' \mid s,a_s)y_j$

$= y_i + \sum_{s\in S_i}[\mathbf{p}_i^*(\mathbf{a})]_s\{r(s,a_s) + \sum_{j=1}^{n}\sum_{s'\in S_j} d(s' \mid s,a_s)y_j\} = y_i +$
$\sum_{s\in S_i}[\mathbf{p}_i^*(\mathbf{a})]_s \bar{c}_i(s,a_s) = y_i + \mathbf{p}_i^*(\mathbf{a})^T \bar{\mathbf{c}}_i(\mathbf{a})$
where $\bar{c}_i(s,a_s) := r(s,a_s) + \sum_{j=1}^{n}\sum_{s'\in S_j} d(s' \mid s,a_s)y_j, \quad s \in S_i$ and
$\bar{\mathbf{c}}_i(\mathbf{a}) := \big(\bar{c}_i(s,a_s)\big)_{s\in S_i}$.

It follows that for each $i = 1,\ldots,n$; (4.4) is the same as:

$$\mathbf{p}_i^*\big(\bar{g}(i)\big)^T \bar{\mathbf{c}}_i\big(\bar{g}(i)\big) > \mathbf{p}_i^*\big(\bar{f}(i)\big)^T \bar{\mathbf{c}}_i\big(\bar{f}(i)\big). \tag{4.5}$$

Since $\mathbf{p}_i^*(\mathbf{a})^T \bar{\mathbf{c}}_i(\mathbf{a})$ is the value of the "strategy" $\mathbf{a}$ $(s \to a_s,\ s \in S_i)$ in the irreducible MDP Γ_i in which the rewards are defined by $\bar{\mathbf{c}}_i$, then the strategy $\bar{g}$ in (4.5) can be computed by one iteration of the simplified policy improvement algorithm.

From the previous results, our Aggregation-Disaggregation Algorithm for solving the limit Markov Control Problem (L) is stated as follows:

Step 1: Select an arbitrary deterministic strategy f in Γ, and set:

$$[f(i)]_s := f(s); \quad s \in S_i; \quad i = 1,2,\ldots n.$$

Step 2: Compute $\mathbf{p}_i^*(f(i))$; $q_{ij}(f(i))$; and $c(i,f(i))$; $\quad i = 1,2,\ldots,n$

$j = 1,2,\ldots,n$. For each $i = 1,\ldots,n$ the computation of $\mathbf{p}_i^*(f(i))$ is done by solving the linear system:

$$\mathbf{x}^{i^T} P_i(f(i)) = \mathbf{x}^i, \quad \sum_{s\in S_i} x_s^i = 1.$$

Step 3: Solve, for the unknowns $\lambda, y_1, y_2, \ldots, y_{n-1}$, the linear system:

$$\lambda + y_i = c(i,f(i)) + \sum_{j=1}^{n} q_{ij}(f(i))y_j \quad i = 1,\ldots,n;\ y_n = 0.$$

Step 4: For each $i = 1,\ldots,n$ compute the deterministic strategy $g(i)$ obtained after one iteration of the simplified policy improvement algorithm for the MDP Γ_i with reward $\bar{\mathbf{c}}_i$ (the starting strategy is $f(i)$).

Step 5: If $g(i) = f(i)$ for all $i = 1,\ldots,n$ STOP.
Otherwise $f(i) \leftarrow g(i)$; $i = 1,2,\ldots,n$ and go to Step 2.

References

[1] R. Aldhaheri and H. Khalil, *Aggregation and optimal control of nearly completely decomposable markov chains*, in Proceedings of the 28th CDC, IEEE, 1989, pp. 1277-1282.

[2] T. R. Bielecki and J. A. Filar, *Singularly Perturbed Markov Control Problem: Limiting Average Cost*, Tech. Rep. 89-04, University of Maryland at Baltimore County, 1989.

[3] M. Cordech, A. Willsky, S. Sastry, and D. Castanon, *Hierarchical aggregation of linear systems with multiple time scales*, IEEE Transactions on Automatic Control, **AC-28** (1983), pp. 1017–1029.

[4] F. Delebecque, *A reduction process for perturbed markov chains*, SIAM Journal of Applied Mathematics, **48** (1983), pp. 325–350.

[5] F. Delebecque and J. Quadrat, *Optimal control of markov chains admitting strong and weak interactions*, Automatica, **17** (1981), pp. 281–296.

[6] E. V. Denardo, *Dynamic Programming*, Prentice-Hall, Eglewood Cliffs, New Jersey, 1982.

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

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

[9] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.

[10] P. Kokotovic, *Application of singular perturbation techniques to control problems*, SIAM Review, **26** (1984), pp. 501–550.

[11] K. G. Murty, *Linear Programming*, Wiley, New York, 1983.

[12] A. A. Pervozvanskii and V. G. Gaitsgori, *Theory of Suboptimal Decisions*, Kluwer, Dordrecht, 1988.

[13] R. G. Phillips and P. Kokotovic, *A singular perturbation approach to modelling and control of markov chains*, IEEE Transactions on Automatic Control, **AC-26** (1981), pp. 1087–1094.

[14] J. Rohlicek and A. Willsky, *Multiple time scale decomposition of discrete time markov chains*, Systems and Control Letters, **11** (1988), pp. 309–314

[15] P. J. Schweitzer, *Perturbation theory and finite markov chains*, Journal of Applied Probability, **5** (1968), pp. 401–413.

[16] P. J. Schweitzer, *Perturbation series expansions for nearly completely-decomposable markov chains*, Teletrafic Analysis and Computer Performance Evaluation, 1986, pp. 319–328.