Math Meth Oper Res (2003) 58:237–245
DOI 10.1007/s001860300290

Mathematical Methods
of Operations Research


# Hierarchical algorithms for discounted and weighted Markov decision processes

**M. Abbad**[1], **C. Daoui**[2]

[1] Faculté des Sciences, B.P. 1014, Rabat, Marokko (e-mail: abbad@fsr.ac.ma)
[2] Faculté des Sciences et Techniques, B.P. 523, Béni-Mellal, Marokko
(e-mail: daouic@Yahoo.com)

Manuscript received: August 2002/Final version received: April 2003

**Abstract.** We consider a discrete time finite Markov decision process (MDP) with the discounted and weighted reward optimality criteria. In [1] the authors considered some decomposition of limiting average MDPs. In this paper, we use an analogous approach for discounted and weighted MDPs. Then, we construct some hierarchical decomposition algorithms for both discounted and weighted MDPs.

**Key words:** Discounted MDP, Weighted MDP, Decomposition, Strongly Connected Classes, Graph theory

## 1 Introduction

Many dynamic planning problems have successfully been analyzed as Markov decision processes; e.g. see [4], [9], [10], and [11]. In these references there are several examples motivating the discounted, average, and weighted reward criteria. First, we consider a discrete time Markov decision process (MDP) with finite state and action spaces under discounted reward optimality criterion, and we propose an algorithm for the computation of an optimal solution which is based on the decomposition by using the technique of levels introduced in [12] for stochastic games. The proposed algorithm finds the optimal value and the corresponding optimal action for any state, step by step, until all states are considered. The computation of an optimal action in any state is done through some restricted MDPs.

The fact that the weighted reward criterion is the weighted sum of a discounted and an average reward criteria, leads to the use of the algorithm above and the algorithm developed in [1] for limiting average MDPs to construct two new algorithms: the first determines $\epsilon$-optimal strategies for the restricted weighted MDPs and the second constructs an $\epsilon$-optimal strategy for the original weighted MDP.

This paper is organized as follows: in Section 2, we define weighted MDPs. In Section 3, we propose a decomposition algorithm to determine a discounted optimal strategy. Finally, in Section 4, we propose a level based algorithm to determine an ultimately deterministic $\epsilon$-optimal strategy for weighted MDPs.

## 2 Definitions and preliminaries

We consider a stochastic dynamic system which is observed at discrete time points $t = 1, 2, \ldots$. At each time point $t$ the state space of the system is denoted by $X_t$ where $X_t$ is a random variable whose values are in a state space $E$. At each time point $t$, if the system is in state $i$, an action $a \in A(i) = \{1, 2, \ldots, m(i)\}$ has to be chosen. In this case, two things happen: a reward $r(i, a)$ is earned immediately, and the system moves to a new state $j$ according to the transition probability $p_{iaj}$. Let $A_t$ be the random variable which represents the action chosen at time $t$.

We denote by $H_t = (E \times A)^{t-1} \times E$ the set of all histories up to time $t$, and by $\Psi = \{(q_1, q_2, \ldots, q_{|A|}) : \sum_{a=1}^{|A|} q_a = 1, q_a \geq 0, 1 \leq a \leq |A|\}$ the set of probability distributions over $A = \bigcup_{i \in E} A(i)$. A strategy $\pi$ is defined by a sequence $\pi = (\pi^1, \pi^2, \ldots)$ where $\pi^t : H_t \to \Psi$ is a decision rule. A Markov strategy is one in which $\pi^t$ depends only on the current state at time $t$. A stationary strategy is a Markov strategy with identical decision rules. A deterministic (or pure) strategy is a stationary strategy whose single decision rule is nonrandomized. An ultimately deterministic strategy is a Markov strategy $\pi = (\pi^1, \pi^2, \ldots)$ such that there exist a deterministic strategy $g$ and an integer $t_0$ such that $\pi^t = g$ for all $t \geq t_0$.

Let $F$, $F_M$, $F_S$, $F_D$ and $F_{UD}$ be the sets of all strategies, Markov strategies, stationary strategies, deterministic strategies, and ultimately deterministic strategies, respectively.

Let $P_\pi(X_t = j, A_t = a \mid X_1 = i)$ be the conditional probability that at time t the system is in state j and the action taken is a, given that the initial state is i and the decision maker uses a strategy $\pi$. Now, if $R_t$ denotes the reward at time $t$, then for any strategy $\pi$ and an initial state $i$, the expectation of $R_t$ is given by $E_\pi(R_t, i) = \sum_{j \in E} \sum_{a \in A(j)} P_\pi(X_t = j, A_t = a \mid X_1 = i) r(j, a)$.

The manner in which the resulting stream of expected rewards $\{E_\pi(R_t, i) : t = 1, 2, \ldots\}$ are aggregated defines the Markov decision processes discussed in the sequel.

In the **discounted reward MDP**, the corresponding overall reward criterion is defined by:

$V_i^\alpha(\pi) = \sum_{t=1}^{\infty} \alpha^{t-1} E_\pi(R_t, i)$, $i \in E$, where $\alpha \in [0, 1)$ is a fixed discount factor. A strategy $f^*$ is called discounted optimal if for all $i \in E$, $V_i^\alpha(f^*) = \max_{\pi \in F} V_i^\alpha(\pi) := V^\alpha(i)$. We will denote this MDP by $\Gamma(\alpha)$.

In the **average reward MDP**, the overall reward criterion is defined by: $\Phi_i(\pi) = \liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} E_\pi(R_t, i)$; $i \in E$. A strategy $f^*$ is called average optimal if for all i $\in E$, $\Phi_i(f^*) = \max_{\pi \in F} \Phi_i(\pi) := V(i)$. We will denote this MDP by $\Gamma$.

In the **weighted reward MDP**, the overall reward criterion is defined by: $\omega_i(\pi) = \lambda(1 - \alpha)\, V_i^\alpha(\pi) + (1 - \lambda)\Phi_i(\pi)$, i $\in E$, where $\lambda \in [0, 1]$ is a fixed weighted parameter, and $\alpha$ is the discount factor in the MDP $\Gamma(\alpha)$. We denote this MDP by $\Gamma(\alpha, \lambda)$. A strategy $f^*$ is called optimal if for all i $\in E$, $\omega_i(f^*) = \max_{\pi \in F} \omega_i(\pi)$. Let $\epsilon > 0$, for any $i \in E$, a strategy $f^*$ is called $\epsilon - i-$optimal if $\omega_i(f^*) \geq \max_{\pi \in F} \omega_i(\pi) - \epsilon$. A strategy $f^*$ is called $\epsilon-$ optimal if $f^*$ is $\epsilon - i-$optimal for all i $\in E$.

**Remark 2.1** *Weighted MDPs were formally introduced in* [7] *even though they can be viewed as special cases of more general models considered in* [3]. *In* [7] *the authors show that optimal strategies may not exist and propose an algorithm to determine an $\epsilon$–optimal strategy.*

## 3 Decomposition of discounted MDP

In this section, we consider discounted MDPs with finite state and action spaces. Let $G = (E, U)$ be the graph associated with the original MDP, that is, the state space represents the set of nodes and $U := \{(i, j) \in E^2 : p_{iaj} > 0$ for some $a \in A(i)\}$ the set of directed arcs. The state space can be partitioned into strongly connected classes $C_1$, $C_2, \ldots$, $C_p$. Note that the strongly connected classes are defined to be the classes with respect to the relation on $G$ defined by: $i$ is strongly connected to $j$ if and only if $i = j$ or there exist a directed path from $i$ to $j$ and a directed path from $j$ to $i$. There are many good algorithms in graph theory for the computation of such partition, e.g., see [6]. Now, we construct by induction the levels of the graph $G$. The level $L_0$ is formed by all classes $C_i$ such that $C_i$ is closed, that is, any arc emanating from $C_i$ has both nodes in $C_i$ . The nth level $L_n$ is formed by all classes $C_i$ such that the end of any arc emanating from $C_i$ is in some level $L_{n-1}, L_{n-2}, \ldots, L_0$.

**Remark 3.1** *Let $C_i$ be a strongly connected class in the level $L_n$ then $C_i$ is closed with respect to the restricted MDP to the state space $E - (L_0 \cup L_1 \cup \ldots \cup L_{n-1})$.*

It is clear that, from Remark 3.1, the following algorithm finds the levels.

**Algorithm 3.1:**
$\Omega \leftarrow E$; n $\leftarrow$ 0; $L_n \leftarrow \{ C_i : C_i$ is closed $\}$
If $L_0 = E$ Stop.
Otherwise, unless $\Omega \neq \emptyset$ do
Delete $L_n$ (i.e $\Omega \leftarrow \Omega - L_n$ and eliminate all arcs coming into $L_n$);
$L_{n+1} \leftarrow \{C_i : C_i$ is closed in the MDP restricted to $\Omega\}$;
$n \leftarrow n + 1$.

In what follows, we construct, by induction, the restricted MDPs corresponding to each level $L_n$, $n = 0, 1, 2, \ldots, L$. Let $(C_{lk})$, $k \in \{1, 2, \ldots, K(l)\}$ be the strongly connected classes corresponding to the nodes in level $l$.

**Construction of the restricted MDPs in level $L_0$:** For each $k = 1, 2, \ldots$, $K(0)$, we denote by $MDP_{0k}$ the restricted MDP corresponding to the class $C_{0k}$ that is the restricted MDP in which the state space is $S_{0k} = C_{0k}$. Note that any restricted MDP, $MDP_{0k}$ is well defined since any class $C_{0k}$ is closed and can be easily solved by a finite algorithm (see [5]).

We denote by $\pi_{0k}$ an optimal strategy and $V^{\alpha}_{0k}(i)$, $i \in C_{0k}$ the optimal value in state $i$.

**Construction of the restricted MDPs in level $L_1$:** For each $k = 1$, $2, \ldots, K(1)$, we denote by $MDP_{1k}$ the restricted MDP defined by:

**State space:** $S_{1k} = C_{1k} \cup \{ j \in L_0 : \exists i \in C_{1k},\ \exists a \in A(i) \text{ and } p_{iaj} > 0 \}$.

**Action spaces:** For each $s \in S_{1k}$, the associated action space is:

$A_{1k}(s) = A(s)$ if $s \in C_{1k}$ and $A_{1k}(s) = \{\theta\}$ if $s \notin C_{1k}$.

**Transition probabilities:** Let $i, j \in S_{1k}$: The associated transition probabilities are:

$p_{1k}(j \mid i, a) = p_{iaj}$ if $i \in C_{1k},\ a \in A(i)$ and $p_{1k}(j \mid i, a) = 1$ if $i = j$ and $i \notin C_{1k}$

**Rewards:** Let $i \in S_{1k}$.

If $i \in C_{1k}$ ; $r_{1k}(i, a) := r(i, a)$.

If $i \notin C_{1k}$ ; $\exists h \in \{1, 2, \ldots, K(0)\} :\ i \in C_{0h}$ and $r_{1k}(i, \theta) := (1 - \alpha) V^{\alpha}_{0h}(i)$.

**Remark 3.2** *The construction of restricted MDPs corresponding to different optimality criteria differs from the definition of rewards. Let $i \in (S_{1k} - C_{1k})$ then there exists $h \in \{1, 2, \ldots, K(0)\}$ such that $i \in C_{0h}$. In order to conserve the optimal value at state i, we define $r_{1k}(i, \theta) := V_{0h}(i)$ and $r_{1k}(i, \theta) := (1 - \alpha) V^{\alpha}_{0h}(i)$ in the case of average MDPs and discounted MDPs respectively.*

**Construction of the restricted MDPs in level $L_n$, $n > 1$:** Let $E_n = \cup\{C_{mk}, m = 0, \ldots, n-1; k = 1, \ldots, K(m)\}$.

Let $V^{\alpha}_{mk}(i)$ be the optimal value in state $i \in E_n$, computed in the previous $MDP_{mk}$ $(m < n)$. For each $k = 1, 2, \ldots, K(n)$, we denote by $MDP_{nk}$ the MDP defined by:

**State space:** $S_{nk} = C_{nk} \cup \{ j \in E_n : p_{iaj} > 0 \text{ for some } i \in C_{nk},\ a \in A(i)\}$.

**Action spaces:** For each $i \in S_{nk}$, the associated action space is $A_{nk}(i) = A(i)$ if $i \in C_{nk}$ and $A_{nk}(i) = \{\theta\}$ if $i \notin C_{nk}$

**Transition probabilities:** For each $i, j \in S_{nk}$; $p_{nk}(j \mid i, a) = p_{iaj}$ if $i \in C_{nk}$, $a \in A(i)$ and $p_{nk}(j \mid i, a) = 1$ if $i = j,\ i \notin C_{nk}$

**Rewards:** Let $i \in S_{nk}$; if $i \in C_{nk}$ then $r_{nk}(i, a) := r(i, a)$.

If $i \notin C_{nk}$ then there exist $m \in \{0, 1, \ldots, n-1\}$ and $h \in \{1, 2, \ldots, K(m)\}$ such that $i \in C_{mh}$ and

$r_{nk}(i, \theta) := (1 - \alpha) V^{\alpha}_{mh}(i)$.

In what follows, we present the main result of this section.

**Theorem 3.1** *Let $V^{\alpha}_{lk}(i)$, $i \in C_{lk}$ be the optimal value in the restricted $MDP_{lk}$, then $V^{\alpha}_{lk}(i)$ is equal to the optimal value $V^{\alpha}(i)$ in the original MDP.*

*Proof* The proof is by induction. For $l = 0$, the result follows from the fact that each $C_{0k}$, $k \in \{1, 2, \ldots, K(0)\}$ is closed. The optimal value $V^{\alpha}$ is the unique solution to [2]:

$$V^{\alpha}(i) = \max_{a \in A(i)} [r(i, a) + \alpha \sum_{j \in C_{0k}} p_{iaj} V^{\alpha}(j)], \quad i \in C_{0k}. \tag{1}$$

The optimal value $V^{\alpha}_{0k}$ is the unique solution to:

$$V^{\alpha}_{0k}(i) = \max_{a \in A_{0k}(i)} [r_{0k}(i, a) + \alpha \sum_{j \in C_{0k}} p_{0k}(j \mid i, a) V^{\alpha}_{0k}(j)], \quad i \in C_{0k}. \tag{2}$$

By using (1), (2), and the fact that $A_{0k}(i) = A(i)$, $r_{0k}(i, a) = r(i, a)$ and $p_{0k}(j \mid i, a) = p_{iaj}$ for all $i \in C_{0k}$, it is clear that $V^{\alpha}_{0k}(i) = V^{\alpha}(i)$ for all $i \in C_{0k}$. Let $n > 0$ and suppose that the result is true for all levels preceding $n$. Now, we shall show that the result is still true for $n$. Let $V^{\alpha}_{nk}(i),\ i \in S_{nk}$ be the optimal value in the restricted $MDP_{nk}$, we have that:

$$V^{\alpha}_{nk}(i) = \max_{a \in A_{nk}(i)} [r_{nk}(i,a) + \alpha \sum_{j \in C_{nk}} p_{nk}(j \mid i,a) V^{\alpha}_{nk}(j) + \alpha \sum_{j \notin C_{nk}} p_{nk}(j \mid i,a) V^{\alpha}_{nk}(j)]. \tag{3}$$

It is clear that from the induction hypothesis, that for all $i \in (S_{nk} - C_{nk})$, $V^{\alpha}_{nk}(i) = V^{\alpha}(i)$ and $V^{\alpha}(i)$ is computed in the preceding levels. Then, for all $i \in C_{nk}$ :

$$V^{\alpha}_{nk}(i) = \max_{a \in A(i)} [r(i,a) + \alpha \sum_{j \in C_{nk}} p_{iaj} V^{\alpha}_{nk}(j) + \alpha \sum_{j \notin C_{nk}} p_{iaj} V^{\alpha}(j)]. \tag{4}$$

Since $V^{\alpha}(i),\ i \in S_{nk}$ is the unique solution to (4) then $V^{\alpha}(i) = V^{\alpha}_{nk}(i)$ for all $i \in C_{nk}$. □

**Corollary 3.1** *Let $\pi_{nk}$ be an optimal deterministic strategy for the restricted $MDP_{nk}$ then for each $i \in C_{nk}$, $\pi_{nk}(i)$ is an optimal action in the original MDP.*

*Proof* For each $i \in C_{nk}$ (from Theorem 3.1) we have that:
$\pi_{nk}(i) = \arg\max_{a \in A_{nk}(i)} \left[r_{nk}(i,a) + \alpha \sum_{j \in S_{nk}} p_{nk}(j \mid i,a) V^{\alpha}_{nk}(j)\right]$
$= \arg max_{a \in A(i)} \left[r(i,a) + \alpha \sum_{j \in S_{nk}} p_{iaj} V^{\alpha}(j)\right].$ □

Now, we propose the following decomposition algorithm for discounted MDPs.

**Algorithm 3.2:**
**Step 1:** Find the strongly connected classes in the graph G.
**Step 2:** Find the levels $L_l,\ l = 0, 1, \ldots, L$ by Algorithm 3.1.
**Step 3:** Find the classes $C_{lk},\ k \in \{1, 2, \ldots, K(l)\}$ belonging to each level.
**Step 4:** For each $l = 0, 1, \ldots, L$ solve the restricted MDPs: $MDP_{lk}$, $k \in \{1, 2, \ldots, K(l)\}$.

**Example 3.1** We consider the original MDP defined by:

**State space:** $E = \{1, 2, \ldots, 6\}$.
**Action spaces:** $A(1) = A(2) = A(3) = A(4) = A(6) = \{1, 2\}; A(5) = \{1\}$.
**Transition probabilities:** $p_{111} = p_{112} = 1/2$; $p_{121} = p_{211} = p_{222} = 1$; $p_{313} = p_{323} = 1$; $p_{412} = 1/3$; $p_{415} = 2/3$; $p_{425} = 1$; $p_{514} = 2/3$; $p_{515} = 1/3$; $p_{615} = 2/3$; $p_{613} = 1/3$; $p_{621} = 1$.
**Rewards:** $r(1,1) = 1$; $r(1,2) = 2$; $r(2,1) = 2$; $r(2,2) = 1$; $r(3,1) = 1$; $r(3,2) = 2$; $r(4,1) = 4$; $r(4,2) = 2$; $r(5,1) = 2$; $r(6,1) = 1$; $r(6,2) = 0$.

*Let* $\alpha = 1/2$. *The steps of Algorithm* 3.2 *are* :
**Step 1:** $C_1 = \{1,2\},\ C_2 = \{3\},\ C_3 = \{4,5\},\ C_4 = \{6\}$.
**Step 2:** $L_0 = C_1 \cup C_2;\ L_1 = C_3;\ L_2 = C_4$.
**Step 4:** In level $L_0$, the state space of the restricted MDP: $MDP_{01}$ is $S_{01} = C_1$, optimal actions are $\pi_{01}(1) = 2$, $\pi_{01}(2) = 1$ and optimal values are $V^{\alpha}_{01}(1) = V^{\alpha}_{01}(2) = 4$. The state space of the restricted MDP: $MDP_{02}$ is $S_{02} = C_2$ and an optimal action is $\pi_{02}(3) = 2$ and the optimal value is $V^{\alpha}_{02}(3) = 4$.

In level $L_1$, the state space of the restricted MDP: $MDP_{11}$ is $S_{11} = C_3 \cup \{2\}$, optimal actions are $\pi_{11}(4) = 1,\ \pi_{11}(5) = 1$ and optimal values are $V_{11}^{\alpha}(4) = 6$, $V_{11}^{\alpha}(5) = 4$.

In level $L_2$, the state space of the restricted MDP: $MDP_{21}$ is $S_{21} = C_4 \cup \{1, 3, 5\}$ and an optimal action is $\pi_{21}(6) = 2$ and the optimal value is $V_{21}^{\alpha}(6) = 3$.

**Remark 3.3** *If the initial state is known, an optimal strategy and the optimal value are computed by solving just few restricted MDPs: one does not need to consider all states. The following algorithm explains this issue when the initial state is i.*

**Algorithm 3.3:**
**Step 1:** Determine the class $C_{mk}$ such that $i \in C_{mk}$.
**Step 2:** Determine the classes $C_{nh}$, $n \in \{0, 1, \ldots, m\}$, $h \in \{1, 2, \ldots, K(n)\}$ such that the end of any arc emanating from $C_{mk}$ is in the classes $C_{nh}$.
**Step 3:** Solve the restricted MDPs: $MDP_{nh}$ found in Step 2.

It is clear that, in the algorithm above, the optimal value and an optimal strategy are obtained by solving only $MDP_{mk}$.

**Remark 3.4** *The results developed in this section for the discounted MDPs can be extended easily to the terminating MDPs:* $\alpha = 1$ *and* $\sum_{j \in E} p_{iaj} < 1$ *for all* $i \in E$, $a \in A(i)$.

## 4 Decomposition of weighted MDPs

In this section, we consider a discrete time Markov Decision Process with finite state and action spaces with the weighted reward criterion. The levels and the restricted MDPs are constructed in similar way as in Section 3.

Now we present the following result which will be used in the rest of this paper.

**Lemma 4.1** *Let* $f_{nk}$ *be an average optimal strategy in the* $MDP_{nk}$ *then there exists an integer N such that* $f_{nk}$ *is* $\epsilon$*-i-optimal in* $\Gamma(\alpha, \alpha^N \lambda)$ *for all* $i \in C_{nk}$.

*Proof* For any $\epsilon > 0$ there exists $N_i$ such that $\alpha^n \lambda (1 - \alpha) V^{\alpha}(i) \leq \epsilon$ wherever $n \geq N_i$. Set $N = \max_{i \in C_{nk}} N_i$, and denote by $\omega[\alpha^N \lambda](\pi) = \alpha^N \lambda (1 - \alpha)$ $V^{\alpha}(\pi) + (1 - \lambda)\Phi(\pi)$ the overall reward with the MDP: $\Gamma(\alpha, \alpha^N \lambda)$. We have for each $i \in C_{nk}$, for any $\pi \in F$, and for any $n \geq N$: $\omega_i[\alpha^n \lambda](\pi) = \alpha^n \lambda (1 - \alpha)\ V_i^{\alpha}(\pi) + (1 - \lambda)\Phi_i\ (\pi) \leq \alpha^n \lambda (1 - \alpha)\ V^{\alpha}(i) + (1 - \lambda) V(i)$. By using the former inequality and the fact that $V(i) = \Phi_{nk}(i, f_{nk}) = V_{nk}(i)$ (see [1]), it is clear that $f_{nk}$ is $\epsilon - i-$optimal for all $i \in C_{nk}$ in $\Gamma(\alpha, \alpha^N \lambda)$. □

In the following we propose an algorithm which constructs an $\epsilon-$ optimal ultimately deterministic strategy for all $i \in C_{nk}$ in $\Gamma(\alpha, \lambda)$.

**Algorithm 4.1:**
**Step 1:** Choose some average optimal strategy $f_{nk}$ in $MDP_{nk}$.

Choose an integer $N = \max_{i \in C_{nk}} N_i$, where $N_i$ is the smallest positive integer such that $\alpha^{N_i}\lambda(1-\alpha)(V^{\alpha}(i) - V_i^{\alpha}(f_{nk})) \leq \epsilon$ ; set $f^{nk} := f_{nk}$.
If $f^{nk}$ is discounted optimal in the $MDP_{nk}$, the algorithm terminates.
**Step 2:** For $h = N$ down to 1.

Select the nonrandomized rule decision $f_{nk}^h$ defined by:
$f_{nk}^h(i) := \arg\max_{a \in A(i)} \left\{ r(i,a)(1-\alpha)\lambda\alpha^h + \sum_{j \in S_{nk}} p_{iaj}\omega_j[\alpha^h\lambda](f^{nk}) \right\}$ for i $\in C_{nk}$.
$f_{nk}^h(i) := \theta$ for $i \in (S_{nk} - C_{nk})$; set $f^{nk} := (f_{nk}^h, f^{nk})$.

**Theorem 4.1** *The ultimately deterministic strategy* $f^{nk} = (f_{nk}^1, f_{nk}^2, \ldots, f_{nk}^N, f_{nk}, f_{nk}, \ldots)$ *constructed by Algorithm* 4.1 *is* $\epsilon-$ *optimal for all* $i \in C_{nk}$ *in* $\Gamma(\alpha, \lambda)$.

*Proof* After Step1, $f^{nk}$ is $\epsilon$-optimal in $\Gamma(\alpha, \alpha^N\lambda)$ by Lemma 4.1. After each iteration in Step 2, $f^{nk}$ is $\epsilon$-optimal in $\Gamma(\alpha, \alpha^{h-1}\lambda)$ by Lemma 3 in [7]. □

**Remark 4.1** *Algorithm 4.1 finds an* $\epsilon-$*optimal strategy for all states belonging to the same strongly connected class* $C_i$ *by solving just the restricted MDP to* $C_i$. *However, in* [7] *for each state* $i \in E$ *an* $\epsilon - i-$*optimal strategy is constructed by solving the whole original MDP.*

**Remark 4.2** *Note that* $f_{nk}$ *and* $f^{nk}$ *refer to a deterministic strategy and an ultimately deterministic strategy respectively.*

In the rest of this section, we will present a new method to construct an $\epsilon-$optimal strategy in $\Gamma(\alpha, \lambda)$ by using the restricted MDPs. To that end, we consider the following lemmata.

**Lemma 4.2** *Let* $f_{nk}$, $n \in \{0, 1, \ldots, L\}$, $k \in \{1, 2, \ldots, K(n)\}$ *be some deterministic strategies in* $MDP_{nk}$ *and define* $f \in F_D$ *such that* $f(i) := f_{nk}(i)$ *for all* $i \in C_{nk}$ *then* $V^{\alpha}(i,f) = V^{\alpha}(i, f_{nk})$ *and* $\Phi_i(f) = \Phi_i(f_{nk})$ *for all* $i \in C_{nk}$.

*Proof* The proof is by induction on $n$. For $n = 0$, $C_{0k}$, $k \in \{1, 2, \ldots, K(0)\}$ are closed, then it is clear that for all $i \in C_{0k}$: $V^{\alpha}(i,f) = V^{\alpha}(i, f_{0k})$ and $\Phi_i(f) = \Phi_i(f_{0k})$.

Suppose that the result is true until the level $n-1$. Now we shall show that the result is still true in the level $n$. Let $i \in C_{nk}$, from the definition of the strategy $f$, it follows that:

$$V^{\alpha}(i,f) = r(i, f_{nk}(i)) + \alpha \sum_{j \in C_{nk}} p_{if_{nk}(i)j} V^{\alpha}(j,f) + \alpha \sum_{j \in (S_{nk} - C_{nk})} p_{if_{nk}(i)j} V^{\alpha}(j,f). \tag{5}$$

It is clear from the induction hypothesis that:

$$V^{\alpha}(i,f_{nk}) = r(i, f_{nk}(i)) + \alpha \sum_{j \in C_{nk}} p_{if_{nk}(i)j} V^{\alpha}(j,f_{nk}) + \alpha \sum_{j \in (S_{nk} - C_{nk})} p_{if_{nk}(i)j} V^{\alpha}(j,f). \tag{6}$$

Since $(V^{\alpha}(i,f_{nk}), i \in C_{nk})$ is the unique solution to the equality above, then $V^{\alpha}(i,f) = V^{\alpha}(i,f_{nk})$ for all $i \in C_{nk}$, $n \in \{0, 1, \ldots, L\}$, $k \in \{1, 2, \ldots, K(n)\}$.

To show the second part, it suffices to use the following classical result: $\lim_{\alpha\to 1^-}(1-\alpha)V^{\alpha}(i,f) = \Phi_i(f)$ for all $i \in C_{nk}$, $f \in F_D$. □

Let $M = \max_{i\in E} N_i$, where $N_i$ is the smallest positive integer such that $\alpha^{N_i}\lambda(1-\alpha)V^{\alpha}(i) \leq \epsilon$.

**Lemma 4.3** *If $f_{nk}$ is average optimal in the $MDP_{nk}$, $n \in \{0,1,\ldots,L\}$, $k \in \{1,2,\ldots,K(n)\}$ then the strategy $f$ constructed above is $\epsilon-$optimal in $\Gamma(\alpha,\alpha^M\lambda)$.*

*Proof* From Lemma 4.2 and definition of $\omega(f)$, we have that $\omega_i[\alpha^p\lambda](f) = \omega_i[\alpha^p\lambda](f_{nk})$ for each $i \in C_{nk}$ and $p \geq 1$. Then, the result follows from Lemma 4.1. □

Now, we suppose that $f_{nk}$, $n \in \{0,1,\ldots,L\}$, $k \in \{1,2,\ldots,K(n)\}$ are average optimal strategies in the $MDP_{nk}$. First, we will construct $\epsilon$-optimal strategies $f^{nk}$ in $\Gamma_{nk}(\alpha,\lambda)$, such that the "tail" of $f^{nk}$ is equal to $f_{nk}$ after stage $M = \max_{i\in E} N_i$ and its "head" is computed with the same manner as in Algorithm 4.1. That is $f^{nk} = (f^1_{nk}, f^2_{nk}, \ldots, f^M_{nk}, f_{nk}, f_{nk}, \ldots.)$ where $f^1_{nk}, f^2_{nk}, \ldots, f^M_{nk}$ are the decision rules computed in Step 2 of Algorithm 4.1. The following theorem constructs an $\epsilon-$optimal strategy in $\Gamma(\alpha,\lambda)$.

**Theorem 4.2** *Let $f = (f^1, f^2, \ldots, f^M, f, f, \ldots) \in F_{UD}$ be defined by: for all $i \in C_{nk}$, $f(i) = f_{nk}(i)$ and $f^h(i) = f^h_{nk}(i)$, $h \in \{1,2,\ldots,M\}$. Then $f$ is $\epsilon-$optimal in $\Gamma(\alpha,\lambda)$.*

*Proof* The result follows from Lemma 4.2 and Theorem 3 in [7]. □

From Theorem 4.2, we can derive the following algorithm.

**Algorithm 4.2:**

**Step 1**: Choose some average optimal strategy $f_A$ as defined in Lemma 4.3. Let $M = \max_{i\in E} N_i$, where $N_i$ is the smallest positive integer such that: $\alpha^{N_i}\lambda(1-\alpha)(V^{\alpha}(i) - V^{\alpha}_i(f_A)) \leq \epsilon$.

If $f_A$ is discounted optimal in the original MDP, the algorithm terminates. Set $f := f_A$.

**Step 2**: For $h = M$ down to 1.

For $n \in \{0,1,\ldots,L\}$, $k \in \{1,2,\ldots,K(n)\}$ select the nonrandomized rule decision $f^h_{nk}$ defined by:

$f^h_{nk}(i) = \arg\max_{a\in A(i)} \left\{ r(i,a)(1-\alpha)\lambda\alpha^h + \sum_{j\in S_{nk}} p_{iaj}\omega_j[\alpha^h\lambda](f) \right\}$ for $i \in C_{nk}$.

$f^h_{nk}(i) = \theta$ for $i \in (S_{nk} - C_{nk})$. Set $f^h(i) = f^h_{nk}(i)$ for $i \in C_{nk}$, $n \in \{0,1,\ldots,L\}$, $k \in \{1,2,\ldots,K(n)\}$, $h \in \{0,1,\ldots,M\}$.

**Remark 4.3** *From Theorem* 4.2, *it follows that the ultimately deterministic policy $f = (f^1, f^2, \ldots, f^M, f_A, f_A, \ldots)$ constructed in Algorithm* 4.2 *above is $\epsilon-$optimal in $\Gamma(\alpha,\lambda)$.*

## References


[1] Abbad M, Boustique H (2003) Decomposition of Limiting Average Markov Decision Problems, to appear in Operations Research Letters

[2] Blackwell D (1962) Discrete Dynamic Programming. Ann. Math. Statist. 33:719–726

[3] Feinberg EA (1982) Controlled Markov Processes with Arbitrary Numerical Criteria. Theo. Prob. Appl. 27:486–503
[4] Feinberg EA, Shwartz A (1994) Markov Decision Models with Weighted Discounted Criteria. Math. Oper. res. 19:152–168
[5] Filar JF, Schultz TA (1988) Communicating MDPs: Equivalence and Properties. Operations Research Letters Vol. 7(6):303–307
[6] Gondran M, Minoux M (1990) Graphes et Algorithmes, 2nd edition
[7] Krass D, Filar JA, Sinha SS (1992) A Weighted Markov Decision Process, Operations Research Vol. 40(6):1180–1187
[8] Krass D (1989) Contributions to the Theory and Applications of Markov Decision Processes, Ph.D. Thesis Johns Hopkins University, Baltimore
[9] Puterman ML (1994) Markov Decision Processes, John Wiley and Sons, Inc., New York
[10] Tijms HC (1986) Stochastic Modeling and Analysis: A computational Approach, John Wiley, New York
[11] White DJ (1985) Real applications of Markov Decision Processes. Interfaces 15(6):73–83
[12] Zeynep M, Avsan, Melike Baykal-Gursoy (1999) A Decomposition Approach for Undiscounted Two Person Zero-Sum Stochastic Games, Mathematical Methods of O.R. Vol. 49(3):483–500