BULETINUL ACADEMIEI DE ŞTIINŢE
A REPUBLICII MOLDOVA. MATEMATICA
Number 1(86), 2018, Pages 34–49
ISSN 1024–7696

An Approach for Determining the Optimal Strategies for an Average Markov Decision Problem with Finite State and Action Spaces

Dmitrii Lozovanu, Stefan Pickl

Abstract. The average reward Markov decision problem with finite state and action spaces is considered and an approach for determining the optimal pure and mixed stationary strategies for this problem is proposed. We show that the considered problem can be formulated in terms of stationary strategies where the objective function is quasi-monotonic (i. e. it is quasi-convex and quasi-concave) on the feasible set of stationary strategies. Using such a quasi-monotonic programming model with linear constraints we ground algorithms for determining the optimal pure and mixed stationary strategies for the average Markov decision problem.

Mathematics subject classification: 90C15, 90C40.
Keywords and phrases: Markov decision processes, Average optimization criterion, Stationary strategies, Optimal strategies.

1 Introduction and problem formulation

In this paper we consider the average reward Markov decision problem for time discrete systems with infinite horizon. The state and action spaces are assumed to be finite. Our aim is to propose an approach for determining the optimal strategies (policies) for an average Markov decision problem that is based on quasimonotonic programming with linear constraints. We show that the considered problem can be formulated in terms of stationary strategies where the objective function is quasi-monotonic (i. e. it is quasi-convex and quasi-concave) on the feasible polyhedron set of stationary strategies. So, we show that our decision problem can be represented as a quasi-monotonic programming problem, and based on such a model we propose algorithms for determining the optimal pure and mixed stationary strategies.

An average Markov decision problem is determined by the following elements:

– a finite set of states X;

– a finite set of actions $A(x)$ for each state $x \in X$;

– a transition probability function $p: X \times \prod_{x\in X} A(x) \times X \rightarrow [0,1]$ that gives the probability transitions $p^a_{x,y}$ from an arbitrary $x \in X$ to an arbitrary $y \in X$ for a fixed action $a \in A(x)$, where $\sum_{y\in X} p^a_{x,y} = 1, \ \forall x \in X, \ a \in A(x)$;

© Dmitrii Lozovanu, Stefan Pickl, 2018

– a step reward $r_{x,a}$ for each state $x \in X$ and every action $a \in A(x)$;

– a starting state $x_0 \in X$.

The control process in the considered problem starts in the state x_0 at the moment of time $t = 0$ where the decision maker selects an action $a_0 \in A(x_o)$ and receives the reward $f(x_0, a_0)$. After that the dynamical system passes randomly to a state $y = x_1 \in X$ according to probability distributions $p^{a_0}_{x_0,y}$, where x_1 is reached at the moment of time $t = 1$. In general, at the moment of time $t \in \{0, 1, 2, \dots\}$ the decision maker observes the state x_t of the system and selects an action $a_t \in A(x_t)$. After that he receiveds the reward $f(x_t, a_t)$ and the system passes randomly to a state $y = x_{t+1}$ according to probability distributions $p^{a_t}_{x_t,y}$. Such a process induces a sequence of rewards $f(x_0, a_0), f(x_1, a_1), \dots, f(x_t, a_t), \dots$ for which the *average reward per transition*

$$\omega_{x_0} = \lim_{t\to\infty} \inf \ \mathsf{E}\left(\frac{1}{t}\sum_{\tau=0}^{t-1} f(x_\tau, a_\tau)\right),$$

is gained. The aim of the decision maker is to determine a strategy of selection the actions in the states that provides the maximal average reward per transition. *A strategy* in a Markov decision problem is a mapping s^i that for every state $x_t \in X$ at the moment of time t provides a probability distribution over the set of actions $A(x_t)$. If the probabilities in these distributions take only values 0 and 1, then s is called a *pure strategy*, otherwise s is called a *mixed strategy.* If these probabilities depend only on the state $x_t = x \in X$ (i. e. s does not depend on t), then s is called a *stationary strategy.*

Let s be a stationary strategy applied by the decision maker. Then such a strategy induces a Markov chain with probability transition matrix $P^s = (p_{x,y^s})$ and the step rewards $f(x, s)$ in the states $x \in X$ that can be determined as follows:

$$p^s_{x,y} = \sum_{a\in A(x)} s_{x,y} p_{x,y}, \ \forall x, y \in X; \quad f(x, s) = \sum_{a\in A(x)} s_{x,a} f(x, a), \forall x \in X.$$

This means that if a stationary strategy s in the control process is applied then the average reward per transition can be calculated by using the formula

$$\omega_{x_0}(s) = \sum_{y\in Y} f(y, s) q^s_{x_0,y},$$

where $q^s_{x,y}$ for $x, y \in X$ represents the elements of the limit matrix $Q^s = (q^s_{x,y})$ for the probability transition matrix P^s induced by the strategy s.

It is well-known (see [8]) that for the average Markov decision problem with finite state and action spaces an optimal strategy always exists and such an optimal strategy can be found in the set of stationary strategies. Moreover, for the considered problem there exists an optimal strategy that corresponds to a pure stationary strategy that is an optimal one for an arbitrary starting state. Taking into account that the set of stationary strategies corresponds to the set of feasible solutions of the following system

$$\begin{cases} \sum\limits_{a\in A(x)} s_{x,a} = 1, & \forall x \in X; \\ \qquad s_{x,a} \geq 0, & \forall x \in X, \ \ \forall a \in A(x), \end{cases} \tag{1}$$

where each extreme point of this set of solutions corresponds to a pure stationary strategy then we can determine the optimal solution by calculating the average reward for each extreme point of system (1) and selecting the best one. Obviously such an approach is not convenient. Suitable algorithms for determining the optimal strategies for the average decision problem based on linear programming and algorithms based on value and policy iteration can be found in [3, 8, 9].

In this paper we propose an approach for determining the optimal strategies that is based on quasimonotonic programming with linear constraints. We show that such an approach allows us to ground new algorithms for determining the optimal solution for the considered problem.

2 Preliminaries

In [8] it is shown that an optimal stationary strategy for an infinite horizon average Markov decision problem with finite state and action spaces can be found by solving the following linear programming problem:
Maximize

$$\varphi(\alpha, \beta) = \sum_{x\in X} \sum_{a\in A(x)} r_{x,a}\alpha_{x,a} \tag{2}$$

subject to

$$\begin{cases} \sum\limits_{a\in A(y)} \alpha_{y,a} - \sum\limits_{x\in X} \sum\limits_{a\in A(x)} p^a_{x,y}\, \alpha_{x,a} = 0, \ \ \forall y \in X; \\ \sum\limits_{a\in A(y)} \alpha_{y,a} + \sum\limits_{a\in A(y)} \beta_{y,a} - \sum\limits_{x\in X} \sum\limits_{a\in A(x)} p^a_{x,y}\beta_{x,a} = \theta_y, \ \forall y \in X; \\ \qquad \alpha_{x,a} \geq 0, \quad \beta_{y,a} \geq 0, \ \ \forall x \in X, \ a \in A(x), \end{cases} \tag{3}$$

where θ_y for $y \in X$ represent arbitrary positive values that satisfy the condition $\sum\limits_{y\in X} \theta_y = 1$, where θ_y for $y \in Y$ are treated as the probabilities of choosing the starting state $y \in X$. In the case $\theta_y = 1$ for $y = x_0$ and $\theta_y = 0$ for $y \in X \setminus \{x_0\}$ we obtain the linear programming model for an average Markov decision problem with fixed starting state x_0.

This linear programming model corresponds to the multichain case of an average Markov decision problem. If each stationary policy in the decision problem induces an ergodic Markov chain then the restrictions (3) can be replaced by the restrictions

$$\begin{cases} \sum\limits_{a\in A(y)} \alpha_{y,a} - \sum\limits_{x\in X} \sum\limits_{a\in A(x)} p_{x,y}^{a}\, \alpha_{x,a} = 0, \quad \forall y \in X; \\ \sum\limits_{y\in X} \sum\limits_{a\in A(y)} \alpha_{y,a} = 1; \\ \qquad\qquad \alpha_{y,a} \geq 0, \quad \forall y \in X, \quad a \in A(y). \end{cases} \tag{4}$$

In the linear programming model (2), (3) the restrictions

$$\sum_{a\in A(y)} \alpha_{y,a} + \sum_{a\in A(y)} \beta_{y,a} - \sum_{x\in X} \sum_{a\in A(x)} p_{x,y}^{a} \beta_{x,a} = \theta_y, \ \forall y \in X$$

with the condition $\sum\limits_{y\in X} \theta_y = 1$ generalize the constraint

$$\sum_{x\in X} \sum_{a\in A(y)} \alpha_{y,a} = 1$$

in the linear programming model (2), (4) for the ergodic case.

The relationship between feasible solutions of problem (2), (3) and stationary strategies in the average Markov decision problem is as follows:

Let (α, β) be a feasible solution of the linear programming problem (2), (3) and denote by $X_\alpha = \{x \in X | \sum\limits_{a\in X} \alpha_{x,a} > 0\}$. Then (α, β) possesses the properties that $\sum\limits_{a\in A(x)} \beta_{x,a} > 0$ for $x \in X \setminus X_\alpha$ and a stationary strategy $s_{x,a}$ that corresponds to (α, β) is determined by

$$s_{x,a} = \begin{cases} \dfrac{\alpha_{x,a}}{\sum\limits_{a\in A(x)} \alpha_{x,a}} & \text{if } \ x \in X_\alpha; \\ \dfrac{\beta_{x,a}}{\sum\limits_{a\in A(x)} \beta_{x,a}} & \text{if } \ x \in X \setminus X_\alpha, \end{cases} \tag{5}$$

where $s_{x,a}$ expresses the probability of choosing the actions $a \in A(x)$ in the states $x \in X$.

Remark 1. Problem (2), (3) can be considered also for the case when $\theta_x = 0$ for some $x \in X$. In particular, if $\theta_x = 0, \ \forall x \in X \setminus \{x_0\}$ and $\theta_{x_0} = 1$ then this problem is transformed into the model with fixed starting state x_0. In this case for a feasible solution (α, β) the subset $X \setminus X_\alpha$ may contain states for which $\sum_{a\in A(x)} \beta_{x,a} = 0$. In such states (5) couldn't be used for determining $s_{d_{\alpha,\beta}(x)}(a)$. Formula (5) can be used for determining the strategies $s_{x,a} = s_{d_{\alpha,\beta}(x)}(a)$ in the states $x \in X$ for which either $\sum_{a\in A(x)} \alpha_{x,a} > 0$ or $\sum_{a\in A(x)} \beta_{x,a} > 0$ and these strategies determine the value of the objective function in the decision problem. In the states $x \in X_0$, where

$$X_0 = \{x \in X | \sum_{a\in A(x)} \alpha_{x,a} = 0, \sum_{a\in A(x)} \beta_{x,a} = 0\},$$

the strategies of selection the actions may be arbitrary because they do not affect the value of the objective function.

As it is shown in [3, 5, 8] for an arbitrary average Markov decision problem it always has an optimal solution that corresponds to a pure stationary strategy. However, the linear programming problem (2), (3) may have a basic solution (α, β) for which the corresponding stationary strategy s determined through (5) is not a pure stationary strategy for the Markov decision problem. So, if we solve the linear programming problem (2), (3) and find an optimal basic solution (α^*, β^*) then the corresponding optimal stationary strategy s^* determined according to (5) may be not a pure strategy. In this paper we formulate a new optimization model in terms of stationary strategies for the average Markov decision problem that allows to determine all optimal pure stationary strategies. The proposed model is related to quasi-monotonic programming in which it is necessary to maximize a quasi-monotonic objective function on a convex polyhedron set.

3 The main results

In this section we present the results that allow us to formulate the average Markov decision problem in terms of stationary strategies as a quasi-monotonic programming problem with linear constraints and to determine the optimal stationary strategies.

3.1 A nonlinear optimization model in terms of stationary strategies for average Markov decision problem

First we show that an average Markov decision problem in terms of stationary strategies can be formulated as follows:
Maximize

$$\psi(s, q, w) = \sum_{x \in X} \sum_{a \in A(x)} f(x, a) s_{x,a} q_x \tag{6}$$

subject to

$$\begin{cases} q_y - \sum\limits_{x \in X} \sum\limits_{a \in A(x)} p_{x,y}^a \, s_{x,a} q_x = 0, & \forall y \in X; \\ q_y + w_y - \sum\limits_{x \in X} \sum\limits_{a \in A(x)} p_{x,y}^a s_{x,a} w_x = \theta_y, & \forall y \in X; \\ \sum\limits_{a \in A(y)} s_{y,a} = 1, & \forall y \in X; \\ s_{x,a} \geq 0, \ \forall x \in X, \ \forall a \in A(x); \quad w_x \geq 0, \ \forall x \in X, \end{cases} \tag{7}$$

where θ_y are the same values as in problem (2), (3) and $s_{x,a}$, q_x, w_x for $x \in X$, $a \in A(x)$ represent the variables that must be found.

Theorem 1. *Optimization problem (6), (7) determines the optimal stationary strategies of the multichain average Markov decision problem.*

Proof. Let us assume that each action set $A(x), x \in X$ contains a single action a'. Then system (3) is transformed into the following system of equations

$$\begin{cases} q_y - \sum\limits_{x \in X} p_{x,y} q_x = 0, & \forall y \in X; \\ q_y + w_y - \sum\limits_{x \in X} p_{x,y} w_x = \theta_y, & \forall y \in X \end{cases}$$

with conditions $q_y, w_y \geq 0$ for $y \in X$, where $q_y = \alpha_{y,a'}$, $w_y = \beta_{y,a'}$, $\forall y \in X$ and $p_{x,y} = p_{x,y}^{a'}$, $\forall x, y \in X$. As it is shown in [8] such a system uniquely determines q_x for $x \in X$ and determines w_x for $x \in X$ up to an additive constant in each recurrent class of $P = (p_{x,y})$. Here q_x represents the limiting probability in the state x when transitions start in the states $y \in X$ with probabilities θ_y and therefore the condition $q_x \geq 0$ for $x \in X$ can be released. Note that w_x for some states may be negative, however always the additive constants in the corresponding recurrent classes can be chosen so that w_x became nonnegative. In general, we can observe that in (7) the condition $w_x \geq 0$ for $x \in X$ can be released and this does not influence the value of objective function of the problem. In the case $|A(x)| = 1$, $\forall x \in X$ the average cost is determined as $\psi = \sum\limits_{x \in X} f(x) q_x$, where $f(x) = f(x, a), \forall x \in X$.

If the action sets $A(x)$, $x \in X$ may contain more than one action then for a given stationary strategy $s \in \overline{S}$ of the selection of the actions in the states we can find the average cost $\psi(s)$ in a similar way as above by considering the probability matrix $P^s = (p_{x,y}^s)$, where

$$p_{x,y}^s = \sum_{a \in A(x)} p_{x,y}^a s_{x,a} \tag{8}$$

expresses the probability transition from a state $x \in X$ to a state $y \in X$ when the strategy s of selections of the actions in the states is applied. This means that we have to solve the following system of equations

$$\begin{cases} q_y - \sum\limits_{x \in X} p_{x,y}^s q_x = 0, & \forall y \in X; \\ q_y + w_y - \sum\limits_{x \in X} p_{x,y}^s w_x = \theta_y, & \forall y \in X. \end{cases}$$

If in this system we take into account (8) then this system can be written as follows

$$\begin{cases} q_y - \sum\limits_{x \in X} \sum\limits_{a \in A(x)} p_{x,y}^a \, s_{x,a} q_x = 0, & \forall y \in X; \\ q_y + w_y - \sum\limits_{x \in X} \sum\limits_{a \in A(x)} p_{x,y}^a s_{x,a} w_x = \theta_y, & \forall y \in X. \end{cases} \tag{9}$$

An arbitrary solution (q, w) of the system of equations (9) uniquely determines q_y for $y \in X$ that allows us to determine the average cost per transition

$$\psi(s) = \sum_{x \in X} \sum_{a \in X} f(x, a) s_{x,a} q_x \tag{10}$$

when the stationary strategy s is applied. If we are seeking for an optimal stationary strategy then we should add to (9) the conditions

$$\sum_{a\in A(x)} s_{x,a}=1,\ \forall x\in X;\ s_{x,a}\geq 0,\ \forall x\in X, a\in A(x) \tag{11}$$

and to maximize (10) under the constraints (9), (11). In such a way we obtain problem (6), (7) without conditions $w_x \geq 0$ for $x\in X$. As we have noted the conditions $w_x\geq 0$ for $x\in X$ do not influence the values of the objective function (6) and therefore we can preserve such conditions that show the relationship of the problem (6), (7) with problem (2), (3). □

Corollary 1. *Let (s,q,w) be a feasible solution of problem (6), (7). Then*

$$\alpha_{x,a}=s_{x,a}q_x,\quad \beta_{x,a}=s_{x,a}w_x,\quad \forall x\in X, a\in A(x) \tag{12}$$

represents a feasible solution (α,β) of problem (2), (3) and $\psi(s,q,w)=\phi(\alpha,\beta)$. If (α,β) is a feasible solution of problem (2), (3) then a feasible solution (s,q,w) of problem (6), (7) can be determined as follows:

$$s_{x,a}=\begin{cases}\dfrac{\alpha_{x,a}}{\sum\limits_{a\in A(x)}\alpha_{x,a}} & \textit{for}\ \ x\in X_\alpha,\ a\in A(x);\\ \dfrac{\beta_{x,a}}{\sum\limits_{a\in A(x)}\beta_{x,a}} & \textit{for}\ \ x\in X\setminus X_\alpha,\ a\in A(x);\end{cases} \tag{13}$$

$$q_x\ =\sum_{a\in A(x)}\alpha_{x,a},\quad w_x\ =\sum_{a\in A(x)}\beta_{x,a}\quad \textit{for}\ \ x\in X.$$

Note that a pure stationary strategy s of problem (6), (7) corresponds to a basic solution (α,β) of problem (2), (3) for which (13) holds, however system (3) may contain basic solutions for which stationary strategies determined through (13) do not correspond to pure stationary strategies. Moreover, two different feasible solutions of problem (2), (3) may generate through (13) the same stationary strategy. Such solutions of system (3) are considered *equivalent solutions* for the decision problem.

Corollary 2. *If (α^i,β^i), $i=\overline{1,k}$, represent the basic solutions of system (3) then the set of solutions*

$$M=\Big\{(\alpha,\beta)|\ (\alpha,\beta)=\sum_{i=1}^{k}\lambda^i(\alpha^i,\beta^i),\ \ \sum_{i=1}^{k}\lambda^i=1,\ \ \lambda^i>0,\ i=\overline{1,k}\Big\}$$

determines all feasible stationary strategies of problem (6), (7) through (13).

An arbitrary solution (α, β) *of system* $\underline{(3)}$ *can be represented as follows:* $\alpha = \sum_{i=1}^{k} \lambda^i \alpha^i$, *where* $\sum_{i=1}^{k} \lambda^i = 1;\ \ \lambda^i \geq 0,\ \ i = \overline{1,k}$, *and* β *represents a solution of the system*

$$\begin{cases} \sum\limits_{a\in A(y)} \beta_{x,a} - \sum\limits_{z\in X}\ \sum\limits_{a\in A(z)} p^a_{z,x}\beta_{z,a} = \theta_x - \sum\limits_{a\in A(x)} \alpha_{x,a},\ \forall x \in X; \\ \\ \qquad \beta_{y,a} \geq 0,\ \ \forall x \in X,\ a \in A(x). \end{cases}$$

If (α, β) *is a feasible solution of problem (2), (3) and* $(\alpha, \beta) \notin M$ *then there exists a solution* $(\alpha', \beta') \in M$ *that is equivalent to* (α, β) *and* $\overline{\psi}(\alpha, \beta) = \overline{\psi}(\alpha', \beta')$.

3.2 A quasi-monotonic programming model in stationary strategies

The main result of the paper that allows us to ground algorithms for determining the optimal pure stationary strategies for the average Markov decision problem is represented by the following theorem.

Theorem 2. *Let an average Markov decision problem be given and consider the function*

$$\psi(s) = \sum_{x\in X}\ \sum_{a\in A(x)} f_{(x,a)} s_{x,a}\, q_x, \tag{14}$$

where q_x *for* $x \in X$ *satisfies the condition*

$$\begin{cases} q_y - \sum\limits_{x\in X}\ \sum\limits_{a\in A(x)} p^a_{x,y}\, s_{x,a} q_x = 0, & \forall y \in X; \\ q_y + w_y - \sum\limits_{x\in X}\ \sum\limits_{a\in A(x)} p^a_{x,y} s_{x,a} w_x = \theta_y, & \forall y \in X. \end{cases} \tag{15}$$

Then on the set S *of solutions of the system*

$$\begin{cases} \sum\limits_{a\in A(x)} s_{x,a} = 1, & \forall x \in X; \\ \qquad s_{x,a} \geq 0, & \forall x \in X,\ a \in A(x) \end{cases} \tag{16}$$

the function $\psi(s)$ *depends only on* $s_{x,a}$ *for* $x \in X$, $a \in A(x)$ *and* $\psi(s)$ *is quasi-monotonic on* S *(i.e.* $\psi(s)$ *is quasi-convex and quasi-concave on* S*).*

Proof. The proof of first part of the theorem is evident because for an arbitrary $s \in \overline{S}$ system (15) uniquely determines q_x for $x \in X$ and determines w_x for $x \in X$ up to a constant in each recurrent class of $P^s = (p^s_{x,y})$.

Let us prove the second part of the theorem.

Assume that $\theta_x > 0, \forall x \in X$ where $\sum_{x\in X} \theta_x = 1$ and consider arbitrary two strategies $s',\ s'' \in \overline{S}$ for which $s' \neq s''$. Then according to Corollary 1 there exist feasible solutions (α', β') and (α'', β'') of the linear programming problem (2), (3) for which

$$\psi(s') = \varphi(\alpha', \beta'),\ \ \psi(s'') = \varphi(\alpha'', \beta''), \tag{17}$$

where

$$\alpha'_{x,a} = s'_{x,a}q'_x, \quad \alpha''_{x,y} = s''_{x,a}q''_x, \quad \forall x \in X, \ a \in A(x);$$
$$\beta'_{x,a} = s'_{x,a}w'_x, \quad \beta''_{x,y} = s''_{x,a}q''_x, \quad \forall x \in X, \ a \in A(x);$$
$$q'_x = \sum_{a\in A(x)} \alpha'_{x,a} \quad w'_{x,a} = \sum_{a\in A(x)} \beta'_{x,a}, \quad \forall x \in X;$$
$$q''_x = \sum_{a\in A(x)} \alpha''_{x,a} \quad w''_{x,a} = \sum_{a\in A(x)} \beta''_{x,a}, \quad \forall x \in X.$$

The function $\varphi(\alpha, \beta))$ is linear and therefore for an arbitrary feasible solution $(\overline{\alpha}, \overline{\beta})$ of problem (2), (3) holds

$$\varphi(\overline{\alpha}, \overline{\beta}) = t\varphi(\alpha', \beta') + (1-t)\varphi(\alpha'', \beta'') \tag{18}$$

if $0 \le t \le 1$ and $(\overline{\alpha}, \overline{\beta}) = t(\alpha', \beta') + (1-t)(\alpha'', \beta'')$.

Note that $(\overline{\alpha}, \overline{\beta})$ corresponds to a stationary strategy $\overline{s}$ for which

$$\psi(\overline{s}) = \varphi(\overline{\alpha}, \overline{\beta}), \tag{19}$$

where

$$\overline{s}_{x,a} = \begin{cases} \dfrac{\overline{\alpha}_{x,a}}{\overline{q}_x} & \text{if } x \in X_{\overline{\alpha}}; \\ \dfrac{\beta_{x,a}}{\overline{w}_x} & \text{if } x \in X \setminus X_{\overline{\alpha}}. \end{cases} \tag{20}$$

Here $X_{\overline{\alpha}} = \{x \in X | \sum_{a\in A(x)} \overline{\alpha}_{x,a} > 0\}$ is the set of recurrent states induced by $P^{\overline{s}} = (p^{\overline{s}}_{x,y})$, where $p^{\overline{s}}_{x,y}$ are calculated according to (8) for $s = \overline{s}$ and

$$\overline{q}_x = tq'_x + (1-t)q'', \quad \overline{w}_x = tw'_x + (1-t)w''_x, \quad \forall x \in X.$$

We can see that $X_{\overline{\alpha}} = X_{\alpha'} \cup X_{\alpha''}$, where $X_{\alpha'} = \{x \in X | \sum_{a\in A(x)} \alpha'_{x,a} > 0\}$ and $X_{\alpha''} = \{x \in X | \sum_{a\in A(x)} \alpha''_{x,a} > 0\}$. The value

$$\psi(\overline{s}) = \sum_{x\in X} \sum_{a\in A(x)} f(x,a)\overline{s}_{x,a}\overline{q}_x$$

is determined by $f(x,a)$, $\overline{s}_{x,a}$ and $\overline{q}_x$ in recurrent states $x \in X_{\overline{\alpha}}$ and it is equal to $\varphi(\overline{\alpha}, \overline{\beta})$. If we use (20) then for $x \in X_{\overline{\alpha}}$ and $a \in A(x)$ we have

$$\overline{s}_{x,a} = \frac{t\alpha'_{x,a} + (1-t)\alpha''_{x,a}}{tq'_x + (1-t)q''_x} = \frac{ts'_{x,a}q'_x + (1-t)s''_{x,a}q''_x}{tq'_x + (1-t)q''_x} =$$

$$= \frac{tq'_x}{tq'_x + (1-t)q''_x}s'_{x,a} + \frac{(1-t)q''_x}{tq'_x + (1-t)q''_x}s''_{x,a}$$

and for $x \in X \setminus X_{\overline{\alpha}}$ and $a \in A(x)$ we have

$$\overline{s}_{x,a} = \frac{t\beta'_{x,a} + (1-t)\beta''_{x,a}}{tw'_x + (1-t)w''_x} = \frac{ts'_{x,a}w'_x + (1-t)s''_{x,a}w''_x}{tw'_x + (1-t)w''_x} =$$

$$= \frac{tw'_x}{tw'_x + (1-t)w''_x}s'_{x,a} + \frac{(1-t)w''_x}{tw'_x + (1-t)w''_x}s''_{x,a}.$$

So, we obtain

$$\overline{s}_{x,a} = t_x s'_{x,a} + (1-t_x)s''_{x,a}, \quad \forall a \in A(x), \tag{21}$$

where

$$t_x = \begin{cases} \dfrac{tq'_x}{tq'_x + (1-t)q''_x} & \text{if } \ x \in X_{\overline{\alpha}}; \\ \dfrac{tw'_x}{tw'_x + (1-t)w''_x} & \text{if } \ x \in X \setminus X_{\overline{\alpha}}. \end{cases} \tag{22}$$

and from (17)–(19) we have

$$\psi(\overline{s}) = t\psi(s') + (1-t)\psi(s''). \tag{23}$$

This means that if we consider the set of strategies

$$S(s', s'') = \{\overline{s} |\ \overline{s}_{x,a} = t_x s'_{x,a} + (1-t_x)s''_{x,a}, \quad \forall x \in X, a \in A(x)\}$$

then for an arbitrary $\overline{s} \in S(s', s'')$ it holds

$$\min\{\psi(s'), \psi(s'')\} \le \psi(\overline{s}) \le \max\{\psi(s'), \psi(s'')\}, \tag{24}$$

i.e $\psi(s)$ is monotone on $S(s', s'')$. Moreover, using (21)–(24) we obtain that $\overline{s}$ possesses the properties

$$\lim_{t\to 1} \overline{s}_{x,a} = s'_{x,a}, \forall x \in X, a \in A(x); \quad \lim_{t\to 0} \overline{s}_{x,a} = s''_{x,a}, \forall x \in X, a \in A(x) \tag{25}$$

and respectively

$$\lim_{t\to 1} \psi(\overline{s}) = \psi(s'); \quad \lim_{t\to 0} \psi(\overline{s}) = \psi(s'').$$

In the following we show that the function $\psi(s)$ is quasi-monotonic on $\overline{S}$. To prove this it is sufficient to show that for an arbitrary $c \in R$ the *sublevel set*

$$L^-_c(\psi) = \{s \in \overline{S} |\ \psi(s) \le c\}$$

and the *superlevel set*

$$L^+_c(\psi) = \{s \in \overline{S} |\ \psi(s) \ge c\}$$

of function $\psi(s)$ are convex. These sets can be obtained respectively from the *sublevel set* $L^-_c(\varphi) = \{(\alpha, \beta) |\ \varphi(\alpha, \beta)) \le c\}$ and the *superlevel set* $L^+_c(\psi) = \{(\alpha, \beta) |\ \varphi(\alpha, \beta)) \ge c\}$ of function $\varphi(\alpha, \beta)$ for the linear programming problem (2), (3) using (13).

Denote by (α^i, β^i), $i = \overline{1,k}$ the basic solutions of system (3). According to Corollary 2 all feasible strategies of problem (2), (3) can be obtained trough (13) using the basic solutions (α^i, β^i), $i = \overline{1,k}$. Each (α^i, β^i), $i = \overline{1,k}$, determines a stationary strategy

$$s^i_{x,a} = \begin{cases} \dfrac{\alpha^i_{x,a}}{q^i_x}, & \text{for } \ x \in X_{\alpha^i}, \ a \in A(x); \\ \dfrac{\beta^i_{x,a}}{w^i_x}, & \text{for } \ x \in X \setminus X_{\alpha^i}, \ a \in A(x) \end{cases} \tag{26}$$

for which $\psi(s^i) = \overline{\psi}(\alpha^i, \beta^i)$ where

$$X_{\alpha^i} = \{x \in X | \sum_{a \in A(x)} \alpha^i_{x,a} > 0\}, \ \ q^i_x = \sum_{a \in A(x)} \alpha^i_{x,a}, \ \ w^i_x = \sum_{a \in A(x)} \beta^i_{x,a}, \ \forall x \in X. \tag{27}$$

An arbitrary feasible solution (α, β) of system (3) determines a stationary strategy

$$s_{x,a} = \begin{cases} \dfrac{\alpha_{x,a}}{q_x}, & \text{for } \ x \in X_{\alpha}, \ a \in A(x); \\ \dfrac{\beta_{x,a}}{w_x}, & \text{for } \ x \in X \setminus X_{\alpha}, \ a \in A(x), \end{cases} \tag{28}$$

for which $\psi(s) = \overline{\psi}(\alpha, \beta)$ where

$$X_{\alpha} = \{x \in X | \sum_{a \in A(x)} \alpha_{x,a} > 0\}, \quad q_x = \sum_{a \in A(x)} \alpha_{x,a}, \ \ w_x = \sum_{a \in A(x)} \beta_{x,a}, \ \forall x \in X.$$

Taking into account that (α, β) can be represented as

$$(\alpha, \beta) = \sum_{i=1}^{k} \lambda^i (\alpha^i, \beta^i), \text{ where } \sum_{i=1}^{k} \lambda^i = 1, \ \ \lambda^i \geq 0, \ i = \overline{1,k}, \tag{29}$$

we have $\overline{\psi}(\alpha, \beta) = \sum\limits_{i=1}^{k} \overline{\psi}(\alpha^i, \beta^i) \lambda^i$ and we can consider

$$X_{\alpha} = \bigcup_{i=1}^{k} X_{\alpha^i}; \quad \alpha = \sum_{i=1}^{k} \lambda^i \alpha^i; \quad q = \sum_{i=1}^{k} \lambda^i q^i; \quad w = \sum_{i=1}^{k} \lambda^i w^i. \tag{30}$$

Using (26)–(30) we obtain:

$$s_{x,a} = \frac{\alpha_{x,a}}{q_x} = \frac{\sum\limits_{i=1}^{k} \lambda^i \alpha^k_{x,a}}{q_x} = \frac{\sum\limits_{i=1}^{k} \lambda^i s^i_{x,a} q^i_x}{q_x} = \sum_{i=1}^{k} \frac{\lambda^i q^i_x}{q_x} s^i_{x,a}, \ \forall x \in X_{\alpha}, \ a \in A(x);$$

$$s_{x,a} = \frac{\beta_{x,a}}{w_x} = \frac{\sum\limits_{i=1}^{k} \lambda^i \beta^k_{x,a}}{w_x} = \frac{\sum\limits_{i=1}^{k} \lambda^i s^i_{x,a} w^i_x}{w_x} = \sum_{i=1}^{k} \frac{\lambda^i w^i_x}{w_x} s^i_{x,a}, \ \forall x \in X \setminus X_{\alpha}, a \in A(x)$$

and

$$q_x = \sum_{i=1}^{k} \lambda^i q_x^i, \qquad w_x = \sum_{i=1}^{k} \lambda^i w_x^i \;\text{ for } \quad x \in X. \tag{31}$$

So,

$$s_{x,a} = \begin{cases} \displaystyle\sum_{i=1}^{k} \frac{\lambda^i q_x^i}{q_x} s_{x,a}^i & \text{if } \; q_x > 0; \\[2ex] \displaystyle\sum_{i=1}^{k} \frac{\lambda^i w_x^i}{w_x} s_{x,a}^i & \text{if } \; q_x = 0, \end{cases} \tag{32}$$

where q_x and w_x are determined according to (31).

We can see that if $\lambda^i, s^i, q^i, \; i = \overline{1,k}$, are given then the strategy s defined by (32) is a feasible strategy because $s_{x,a} \geq 0, \forall x \in X, a \in A(x)$ and $\sum_{a \in A(x)} s_{x,a} = 1, \; \forall x \in X$. Moreover, we can observe that $q_x = \sum\limits_{i=1}^{k} \lambda^i q_x^i, \;\; w_x = \sum\limits_{i=1}^{k} \lambda^i w_x^i$ for $x \in X$ represent a solution of system (15) for the strategy s defined by (32). This can be verified by introducing (31) and (32) in (15); after such a substitution all equations from (15) are transformed into identities. For $\psi(s)$ we have

$$\psi(s) = \sum_{x \in X} \sum_{a \in A(x)} f(x,a) s_{x,a} q_x = \sum_{x \in X_\alpha} \sum_{a \in A(x)} f(x,a) \sum_{i=1}^{k} \left(\frac{\lambda^i q_x^i}{q_x} s_{x,a}^i \right) q_x =$$

$$\sum_{i=1}^{k} \Bigg(\sum_{x \in X_{\alpha^i}} \sum_{a \in A(x)} f(x,a) s_{x,a}^i q_x^i \Bigg) \lambda^i = \sum_{i=1}^{k} \psi(s^i) \lambda^i,$$

i. e.

$$\psi(s) = \sum_{i=1}^{k} \psi(s^i) \lambda^i, \tag{33}$$

where s is the strategy that corresponds to (α, β).

Thus, assuming that the strategies $s^1, s^2, \ldots, s^k$ correspond to basic solutions $(\alpha^1, \beta^1), (\alpha^2, \beta^2), \ldots, (\alpha^k, \beta^k)$ of problem (2), (3) and $s \in \overline{S}$ corresponds to an arbitrary solution (α, β) of this problem that can be expressed as convex combination of basic solutions of problem (2), (3) with the corresponding coefficients $\lambda^1, \lambda^2, \ldots, \lambda^k$, we can express the strategy s and the corresponding value $\psi(s)$ by (31)–(33). In general the representation (31)–(33) of strategy s and of the value $\psi(s)$ is valid for an arbitrary finite set of strategies from $\overline{S}$ if (α, β) can be represented as convex combination of the finite number of feasible solutions $(\alpha^1, \beta^1), (\alpha^2, \beta^2), \ldots, (\alpha^k, \beta^k)$ that correspond to $s^1, s^2, \ldots, s^k$; in the case $k = 2$ from (31)–(33) we obtain (21)–(23). It is evident that for a feasible strategy $s \in S$ the representation (31), (32) may be not unique, i.e. two different vectors $\overline{\Lambda} = (\overline{\lambda}^1, \overline{\lambda}^2, \ldots, \overline{\lambda}^k)$ and $\overline{\overline{\Lambda}} = \overline{\overline{\lambda}}^1, \overline{\overline{\lambda}}^2, \ldots, \overline{\overline{\lambda}}^k$ may be that determine the same strategy s via (31), (32). If $s^1, s^2, \ldots, s^k$ represent the system of linear independent basic solutions of system (16) then an arbitrary strategy $s \in \overline{S}$ is determined according to (31), (32) where $\lambda^1, \lambda^2, \dots, \lambda^k$ correspond to a solution of the following system

$$\sum_{i=1}^{k} \lambda^i = 1; \quad \lambda^i \geq 0, \quad i = \overline{1,k}.$$

Consequently, the sublevel set $L_c^-(\psi)$ of function $\psi(s)$ represents the set of strategies s determined by (31), (32), where $\lambda^1, \lambda^2, \dots, \lambda^k$ satisfy the condition

$$\begin{cases} \sum\limits_{i=1}^{k} \psi(s^i)\lambda^i \leq c; \\ \sum\limits_{i=1}^{k} \lambda^i = 1; \quad \lambda^i \geq 0, \quad i = \overline{1,k}, \end{cases} \tag{34}$$

and the superlevel set $L_c^+(\psi)$ of $\psi(s)$ represents the set of strategies s determined by (31), (32), where $\lambda^1, \lambda^2, \dots, \lambda^k$ satisfy the condition

$$\begin{cases} \sum\limits_{i=1}^{k} \psi(s^i)\lambda^i \geq c; \\ \sum\limits_{i=1}^{k} \lambda^i = 1; \quad \lambda^i \geq 0, \quad i = \overline{1,k}. \end{cases} \tag{35}$$

Respectively the level set $L_c(\psi) = \{s \in \overline{S} |\ \psi(s) = c\}$ of function $\psi(s)$ represents the set of strategies s determined by (31), (32), where $\lambda^1, \lambda^2, \dots, \lambda^k$ satisfy the condition

$$\begin{cases} \sum\limits_{i=1}^{k} \psi(s^i)\lambda^i = c; \\ \sum\limits_{i=1}^{k} \lambda^i = 1; \quad \lambda^i \geq 0, \quad i = \overline{1,k}. \end{cases} \tag{36}$$

Let us show that $L_c^-(\psi)$, $L_c^+(\psi)$, $L_c(\psi)$ are convex sets. We present the proof of convexity of sublevel set $L_c^-(\psi)$. The proof of convexity of $L_c^+(\psi)$ and $L_c(\psi)$ is similar to the proof of convexity of $L_c^-(\psi)$.

Denote by Λ the set of solutions $(\lambda^1, \lambda^2, \dots, \lambda^k)$ of system (34). Then from (31), (32), (34) we have $L_c^-(\psi) = \prod\limits_{x \in X} \hat{S}_x$ where $\hat{S}_x$ represents the set of strategies

$$s_{x,a} = \begin{cases} \dfrac{\sum_{i=1}^{k} \lambda^i q_x^i s_{x,a}^i}{\sum_{i=1}^{k} \lambda^i q_x^i} & \text{if } \sum_{i=1}^{k} \lambda^i q_x^i > 0, \\ \dfrac{\sum_{i=1}^{k} \lambda^i w_x^i s_{x,a}^i}{\sum_{i=1}^{k} \lambda^i w_x^i} & \text{if } \sum_{i=1}^{k} \lambda^i q_x^i = 0, \end{cases} \qquad a \in A(x)$$

in the state $x \in X$ determined by $(\lambda^1, \lambda^2, \dots, \lambda^k) \in \Lambda$.

For an arbitrary $x \in X$ the set Λ can be represented as follows $\Lambda = \Lambda_x^+ \cup \Lambda_x^0$, where

$$\Lambda_x^+ = \{(\lambda^1, \lambda^2, \dots, \lambda^k) \in \Lambda | \sum_{i=1}^{k} \lambda^i q_x^i > 0\}, \ \ \Lambda_x^0 = \{(\lambda^1, \lambda^2, \dots, \lambda^k) \in \Lambda | \sum_{i=1}^{k} \lambda^i q_x^i = 0\}$$

and $\sum_{i=1}^k \lambda^i w_x^i > 0$ if $\sum_{i=1}^k \lambda^i q_x^i = 0$. Therefore $\hat{S}_x$ can be expressed as follows $\hat{S}_x = \hat{S}_x^+ \cup \hat{S}_x^0$, where $\hat{S}_x^+$ represents the set of strategies

$$s_{x,a} = \frac{\sum_{i=1}^k \lambda^i q_x^i s_{x,a}^i}{\sum_{i=1}^k \lambda^i q_x^i}, \text{ for } a \in A(x) \tag{37}$$

in the state $x \in X$ determined by $(\lambda^1, \lambda^2, \dots, \lambda^k) \in \Lambda_x^+$ and $\hat{S}_x^0$ represents the set of strategies

$$s_{x,a} = \frac{\sum_{i=1}^k \lambda^i w_x^i s_{x,a}^i}{\sum_{i=1}^k \lambda^i w_x^i}, \text{ for } a \in A(x) \tag{38}$$

in the state $x \in X$ determined by $(\lambda^1, \lambda^2, \dots, \lambda^k) \in \Lambda_x^0$.

Thus, if we analyze (37) then observe that $s_{x,a}$, for a given $x \in X$, represents a linear-fractional function with respect to $\lambda^1, \lambda^2, \dots, \lambda^k$ defined on a convex set Λ_x^+ and $\hat{S}_x^+$ is the image of $s_{x,a}$ on Λ_x^+. Therefore $\hat{S}_x^+$ is a convex set. If we analyze (38) then observe that $s_{x,a}$, for given $x \in X$, represents a linear-fractional function with respect to $\lambda^1, \lambda^2, \dots, \lambda^k$ on the convex set Λ_x^0 and $\hat{S}_x^0$ is the image of $s_{x,a}$ on Λ_x^0. Therefore $\hat{S}_x^0$ is a convex set (see [1]). Additionally, we can observe that $\Lambda_x^+ \cap \Lambda_x^0 = \emptyset$ and in the case $\Lambda_x^+, \Lambda_x^0, \neq \emptyset$ the set Λ_x^0 represents the limit inferior of Λ_x^+. Using this property and taking into account (25) we can conclude that each strategy $s_x \in \hat{S}_x^0$ can be regarded as the limit of a sequence of strategies $\{s_x^t\}$ from $\hat{S}_x^+$. Therefore we obtain that $\hat{S}_x = \hat{S}_x^+ \cup \hat{S}_x^0$ is a convex set. This involves the convexity of the sublevel set $L_c^-(\psi)$. In an analogues way using (35) and (36) we can show that the superlevel set $L_c^+(\psi)$ and the level set $L_c(\psi)$ are convex sets. This means that the function $\psi(s)$ is quasi-monotonic on $\overline{S}$. So, if $\theta_x > 0, \forall x \in X$ and $\sum_{x \in X} \theta_x = 1$ then the theorem holds.

If $\theta_x = 0$ for some $x \in X$ then the set $X \setminus X_\alpha$ may contain states for which $\sum_{a \in A(x)} \alpha_{x,a} = 0$ and $\sum_{a \in A(x)} \beta_{x,a} = 0$ (see Remark 1 and Lemma 1). In this case X can be represented as follows: $X = (X \setminus X_0)) \cup X_0$, where $X_0 = \{x \in X | \sum_{a \in A(x)} \alpha_{x,a} = 0; \sum_{a \in A(x)} \beta_{x,a} = 0\}$. For $x \in X \setminus X_0$ the convexity of $\hat{S}_x$ can be proved in the same way as for the case $\theta_x > 0, \forall x \in X$. If $X_0 \neq \emptyset$ then for $x \in X_0$ we have $\hat{S}_x = \overline{S}_x$ and the convexity of $\hat{S}_x$ is evident. So, the theorem holds. □

Similar results can be extended for Markov decision problems with discounted reward criterion for which the problem of determining the optimal stationary strategies can be formulated as quasi-monotonic programming models with linear constraints. Such models have been considered in [4–6].

3.3 Algorithms based on quasi-monotonic programming

Based on Theorem 2, we can determine an optimal stationary strategy using classical descent methods for the maximization of quasi-monotonic function (14) on a convex polyhedron set S (see [1, 2]). In particular if we are seeking for a pure optimal stationary strategy then we can apply the following iterative procedure: Fix an arbitrary pure strategy s^0 that is a basic solution of system (16), find a solution (q^0, w^0) of system (15) with respect to q_x and w_x and calculate $\omega(s^0) = \sum_{x\in X}\sum_{a\in A(x)} f_{(x,a)} s^0_{x,a}\, q^0_x$ (here q^0 is determined uniquely from (15) for a given s^0). Then find a "neighbour" basic solution s^1 for s^0 in S, determine a solution (q^1, w^1) of system (15) and calculate $\omega(s^1) = \sum_{x\in X}\sum_{a\in A(x)} f_{(x,a)} s^1_{x,a}\, q^1_x$. If for an arbitrary "neighbour" basic solution s^1 for s^0 it holds $\omega(s^0) \geq \omega(s^1)$ then s^0 is an optimal pure stationary strategy; otherwise we find a "neighbour" basic solution s^2 for s^1 and in a similar way calculate $\omega(s^2) = \sum_{x\in X}\sum_{a\in A(x)} f_{(x,a)} s^2_{x,a}\, q^2_x$. If for an arbitrary "neighbour" basic solution s^2 for s^1 it holds $\omega(s^1) \geq \omega(s^2)$ then s^1 is an optimal pure stationary strategy; otherwise we find a "neighbour" basic solution s^3 for s^2 and so on. In a finite number of steps we determine an optimal basic solution s^k of system (16) that corresponds to a pure stationary strategy for the average Markov decision problem.

It is easy to observe that the convergence of some algorithms for determining the optimal stationary strategies from [3, 5, 8, 9] can be grounded using the proposed optimization models and Theorem 1. Additionally, the proposed model can be useful for studying the average stochastic positional games in pure stationary strategies [7].

4 Conclusion

An average Markov decision problem with finite state and action spaces can be formulated and studied in terms of stationary strategies using optimization models (6), (7) and (14)–(16). Classical optimization methods and the corresponding algorithms for the maximization of a quasi-linear function (14), (15) on the convex polyhedron set determined by (16) can be applied for finding the optimal pure stationary strategies in the average Markov decision problem.

References

[1] Boyd S., Vandenberghe L. *Convex Optimization.* Cambridge University Press, Cambridge, 2004.

[2] Kruk S., Wolkowicz H. *Pseudolinear programming.* SIAM Review, 1999, No. 41(4), 795–805.

[3] Hu Q., Yue W. *Markov Decision Processes with their Applications.* Springer, New York, 2008.

[4] Lozovanu D. *The game-theoretical approach to Markov decision problems and determining Nash equilibria for stochastic positional games.* Int. J. of Mathematical Modelling and Numercal Optimization, 2011, No. 2(2), 154–158.

[5] Lozovanu D., Pickl S. *Optimization of Stochastic Discrete Systems and Control on Complex Networks.* Springer, 2015.

[6] Lozovanu D., Pickl S. *Determining the optimal strategies for discrete control problems on stochastic networks with discounted costs.* Discrete Applied Mathematics, 2015, **182**, 169–180.

[7] Lozovanu D., Pickl S. *On Nash equilibria for stochastic games and determining the optimal strategies of the players.* Contribution to game theory and management, St. Petersburg University, 2015, **VIII**, 187–198.

[8] Puterman M. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley, New Jersey 2005.

[9] White D. *Markov Decision Processes.* Wiley, New York, 1993.

Received December 27, 2017

Dmitrii Lozovanu
Institute of Mathematics and Computer Science
5 Academiei str., Chişinău, MD−2028
Moldova
E-mail:*lozovanu@math.md*

Stefan Pickl
Institute for Theoretical Computer Science
Mathematics and Operations Research
Universität der Bundeswehr
München, 85577 Neubiberg-München
Germany,
E-mail:*stefan.pickl@unibw.de*