SENSITIVE DISCOUNT OPTIMALITY VIA NESTED LINEAR PROGRAMS FOR ERGODIC MARKOV DECISION PROCESSES

K.E. Avrachenkov

CIAM,
School of Mathematics,
University of South Australia,
The Levels, SA 5095, Australia,
e-mails: K.Avrachenkov@unisa.edu.au

E. Altman

INRIA,
2004 Route des Lucioles,
B. P. 93, 06902,
Sophia-Antipolis Cedex, France,
e-mail: Eitan.Altman@sophia.inria.fr

ABSTRACT

In this paper we discuss the sensitive discount optimality for Markov decision processes. The n-discount optimality is a refined selective criterion, that is a generalization of the average optimality and the bias optimality. Our approach is based on the system of *nested* linear programs. In the last section we provide an algorithm for the computation of the Blackwell optimal policy. The n-discount optimal policies are obtained as by-product of this algorithm. Here we restrict ourselves to the case of completely ergodic Markov decision processes.

Key Words: Markov decision processes, sensitive discount optimality, Blackwell optimality, nested linear programs.

1. INTRODUCTION

Let us consider a Markov decision process (MDP) with the finite state space $S = \{1, ..., N\}$ and transition probabilities $p_{iaj} = P\{X_{t+1} = j | X_t = i, Y_t = a\}$, where action a is taken from the action set $A(i)$. In this section we discuss briefly the main properties of MDP. An interested reader can find a more detailed study of MDP in the folowing books [3, 7, 11] and references therein.

One gains the reward r_{ia} whenever action $a \in A(i)$ is chosen in state $i \in S$. According to [3, 11] we can consider only the class of stationary policies C_s. Every stationary policy $\pi \in C_s$ can be represented by probability distribution $\pi_{ia} = P\{Y_t = a | X_t = i\}$. Moreover, we deal often only with deterministic policies $f \in C_d$, which can be defined by the function $f(i) = a, a \in A(i)$.

For any stationary policy $\pi \in C_s$ we can define the corresponding transition matrix and the reward vector

$$P(\pi) := \sum_{a \in A(i)} p_{iaj}\pi_{ia}, \quad r(\pi) := \sum_{a \in A(i)} r_{ia}\pi_{ia}.$$

The *expected average reward* $g_i(\pi)$ and the *expected discounted reward* $v_i^\lambda(\pi)$ can be expressed as follows:

$$g_i(\pi) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left[P^{t-1}(\pi) r(\pi)\right]_i$$

and

$$v_i^\lambda(\pi) := \sum_{t=1}^{\infty} \lambda^{t-1} \left[P^{t-1}(\pi) r(\pi)\right]_i,$$

where $i \in S$ is an initial state and $\lambda \in (0, 1)$ is a discount factor. One can use the interest rate $\rho = (1 - \lambda)/\lambda \in (0, \infty]$ instead of the discount factor λ. Note that $\lambda \uparrow 1$ is equivalent to $\rho \downarrow 0$. Then the expected discount reward $v^\rho(\pi) := v^{\lambda(\rho)}(\pi)$ can be expanded as a Laurent series [1, 10, 11] for sufficiently small ρ

$$v^\rho(\pi) = (1 + \rho) \left[\rho^{-1} y_{-1}(\pi) + \sum_{n=0}^{\infty} \rho^n y_n(\pi)\right], \quad (1)$$

where $y_{-1}(\pi) = P^*(\pi) r(\pi)$, $y_0 = H(\pi) r(\pi)$, and $y_n = (-1)^n H(\pi)^{n+1} r(\pi)$, $n = 1, 2, ...$ The *ergodic projection* $P^*(\pi)$ and the *reduced resolvent* (or *deviation matrix*) $H(\pi)$ are defined by

$$P^*(\pi) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P^{t-1}(\pi),$$

$$H(\pi) := (I - P(\pi) + P^*(\pi))^{-1} - P^*(\pi).$$

The next properties of these matrices [11] will be used extensively.

$$P^*(\pi)P(\pi) = P(\pi)P^*(\pi) = P^*(\pi) \quad (2)$$

Part of this work was done while the second author was visiting the University of South Australia and was supported by the Australian Research Council under Grant A49532260.

$$P^*(\pi)H(\pi) = H(\pi)P^*(\pi) = 0 \tag{3}$$

$$(I - P(\pi))H(\pi) = H(\pi)(I - P(\pi)) = I - P^*(\pi) \tag{4}$$

Note that in particular $y_{-1}(\pi) = g(\pi)$ and $y_0 = h(\pi)$, where $g(\pi)$ is the expected average reward vector defined above and $h(\pi)$ is so-called the expected bias reward vector.

Let us now discuss several optimality criteria [7, 11, 12]. The discount optimality and the average optimality are basic criteria in MDP models.

Definition 1.1 *The stationary policy π_* is called the discount optimal for fixed $\lambda \in (0,1)$ if $v_i^\lambda(\pi_*) \geq v_i^\lambda(\pi)$ for each $i \in S$ and all $\pi \in C_s$.*

Definition 1.2 *The stationary policy π_* is called the average optimal if $g_i(\pi_*) \geq g_i(\pi)$ for each $i \in S$ and all $\pi \in C_s$.*

The Laurent expansion (1) allows us to define more selective optimality criteria [11, 12].

Definition 1.3 *We say that a policy $\pi^* \in C_s$ is n-discount optimal for some integer $n \geq 0$ if*

$$[y_n(\pi^*)]_i \geq [y_n(\pi)]_i$$

for each $i \in S$ and all policies π that are $(n-1)$-discount optimal. By convention, the average optimal policy is (-1)-discount optimal policy.

In particular, 0-discount optimal policy is called *bias optimal*.

It is well known that there exists a stationary (and even deterministic) policy that is n-discount optimal for all $n \geq -1$ [10, 11, 12]. It is referred to as the *Blackwell optimal* policy and defined in the following way.

Definition 1.4 *We say that a policy π_* is Blackwell optimal if there exists some $\rho_0 > 0$ such that for all $\rho \in (0, \rho_0]$ $v^\rho(\pi^*) \geq v^\rho(\pi)$ for all $\pi \in C_s$.*

In other words, Blackwell optimal policy is the policy which is discount optimal policy for any discount factor sufficiently close to one. The following important result says that in order to find the Blackwell optimal policy it is sufficient to find n-discount optimal policy for some finite n [9, 13].

Theorem 1.1 *Suppose π_* is $(|S| - r - 1)$-discount optimal, where r is the number of ergodic classes in the Markov chain induced by transition matrix $P(\pi_*)$. Then π_* is Blackwell optimal. And moreover π_* is n-discount optimal for all $n \geq -1$.*

The bias optimality and Blackwell optimality were first introduced in the pioneering work of Blackwell [1]. Later Veinott defined n-discount optimality [12] that provides a unified treatment of average optimality, bias optimality as well as Blackwell optimality. Moreover, in [12] he provided an idea of the algorithm for the computation of n-discount optimal policy based on lexicographic ordering. Veinott developed further his algorithm in [13], where the proof of Theorem 1.1 can be also found. Recently Puterman [11] refined this lexicographic method. An interesting approach to the computation of Blackwell optimal policy was proposed by Hordijk *et. al.* [5]. This approach is based on asymptotic linear programming [6]. Unfortunately, the method of [5] cannot be used for the calculation of n-discount optimal policies that are not Blackwell optimal. In [2] Denardo proposed an algorithm for determining the bias optimal policy via nested linear programs. This algorithm was later refined by Kallenberg [7, 8]. Our algorithm for determining the n-discount optimal policy and Blackwell optimal policy is a generalization of the latter approach. We would like to note that Denardo [2] has already pointed out the possibility of such extension. Since our approach is based on *truncated* MDP models, the number of variables in associated nested linear programms decreases successively. The algorithm can be used for the computation of Blackwell optimal as well as n-discount optimal policies. Moreover, since it is based on linear programming, the algorithm can be easily extended to the MDP model with general Borel state space [4, 14].

In this paper we restrict ourselves only to the case of completely ergodic MDP model. That is,

Assumption 1.1 *For any stationary policy π the Markov chain induced by transition matrix $P(\pi)$ has single ergodic class and there is no transient states.*

The above assumption implies that the ergodic projection $P^*(\pi)$ associated with any stationary policy π has identical rows. The latter implies that the expected average reward vector $g(\pi) = P^*(\pi)r(\pi)$ has identical elements as well. Therefore we shall use the notation $g(\pi)$ not only for the average reward vector but also for any its component. Since in every particular case it is clear whether we use a vector or a scalar, this notation should not cause any confusion.

The reminder of the paper is organized as follows: In the next two sections we briefly review the known results on average optimality and bias optimality [2, 7]. In Section 4 we present our general approach to n-discount optimality. The formal algorithm for the computation of Blackwell optimal as well as n-discount optimal policies via nested linear programs is given in Section 5. In Conclusion section we point out the possible direction of future research.

2. THE AVERAGE OPTIMALITY

In this section we recall the well-known algorithm [3, 7, 11] for finding an average optimal policy for the completely ergodic Markov decision process (see Assumption 1.1). This algorithm is based on the pair of dual linear programs. The primal program is

$$\min\ \tilde{g} \tag{5}$$

$$\text{s.t.} \qquad \tilde{g} + \sum_j (\delta_{ij} - p_{iaj})\tilde{h}_j \geq r_{ia}, i \in S, a \in A(i)\}$$

and the corresponding dual program is

$$\max \sum_{i,a} r_{ia} x_{ia} \tag{6}$$

$$\text{s.t.} \quad \sum_{i,a} (\delta_{ij} - p_{iaj}) x_{ia} = 0, \quad \sum_{i,a} x_{ia} = 1, \quad x_{ia} \geq 0$$

Let (g_*, h_*) be an optimal solution of the linear program (5), then $g_* = g(\pi_*) = \max\{g(\pi) | \pi \in C_s\}$, where π_* is an average optimal policy. The average optimal policy can be obtained from the dual linear program (6). Suppose that x_* is a basic optimal solution, then the deterministic average optimal policy f can be obtained as follows

$$f(i) = a_i, \quad \text{such that} \quad x_{*ia_i} > 0. \tag{7}$$

Note that the above procedure defines a one-to-one correspondence between the sets of all basic optimal solutions of (6) and the set of all average optimal deterministic policies [3, 11].

3. THE BIAS OPTIMALITY

Now let us discuss the approach for the computation of the bias optimal policy based on the system of two nested linear programs [2, 7]. The exposition of this section follows closely the book of Kallenberg [7].

We first solve the pair of dual linear programs (5) and (6) for the computation of the stationary average optimal policy. As a result, we have the optimal solutions (g_*, h_*) and x_* of the primal and dual linear programs respectively. Recall that g_* is an optimal value of the average reward. Let us define the following set.

$$A_0(i) \stackrel{def}{=} \{a \in A(i) \mid g_* + \sum_j (\delta_{ij} - p_{iaj}) h_{*j} = r_{ia}\} \tag{8}$$

Remark 3.1 *The above set can be determined in two ways: either by a simple search procedure among all actions or from the set of basic optimal solutions of the dual problem* (6)*, namely, if x_* is a basic optimal solution, then $a \in A_0(i)$ whenever $x_{*ia} > 0$. The second approach will be justified by the next proposition.*

Let us remind that all statements of the present paper are made under Assumption 1.1 .

Proposition 3.1 *Let f be any deterministic and stationary average optimal policy. Then,*
(i) $f(i) \in A_0(i),\ i \in S$, and
(ii) the corresponding bias vector can be calculated by the formula

$$h(f) = h_* - P^*(f) h_*.$$

PROOF: See [7].

According to Definition 1.3 , the stationary policy d is bias optimal or 0-discount optimal if $h_i(d) = \max_f\{h_i(f) | g(f) = g_*\}$. In other words the average optimal policy, which supplies the optimal value to the bias vector, is the bias optimal policy. It follows from the second part of the Proposition 3.1 that the maximization of $h(f) = h_* - P^*(f)h_*$ is equivalent to the maximization of $-P^*(f)h_*$ over all average optimal policies f.

Note that $g^{(0)} := -P^*(f)h_*$ is the expected average reward vector for the Markov decision process with transition matrix $P(f)$ and immediate reward vector $r^{(0)}_{ia} := -h_{*i}$. Thus we can consider another Markov decision process with state space S and *truncated* action set $A_0(i)$. Note that the part (i) of Proposition 3.1 guarantees that $A_0(i) \neq \emptyset$ for every state $i \in S$. Moreover, according to Assumption 1.1, all stationary policies induce the Markov chain with single ergodic class (without transient states). Hence all policies constructed from the truncated action sets $A_0(i),\ i \in S$ also induce an irreducible Markov chain. The latter implies that the new *truncated* MDP model is also completely ergodic and we can use the linear programming approach from Section 2 again.

Let us now present the dual and primal linear programs for the truncated Markov decision process. The primal program is

$$\min\ \tilde{g}^{(0)} \tag{9}$$

$$\text{s.t.} \quad \tilde{g}^{(0)} + \sum_j (\delta_{ij} - p_{iaj})\, \tilde{h}^{(0)}_j \geq -h_{*i}, i \in S,\ a \in A_0(i).$$

The optimal solution of this program $g^{(0)}_*$ is the optimal average reward for the truncated model. The dual program is of the form

$$\max \sum_i (-h_{*i}) \sum_{a \in A_0(i)} x^{(0)}_{ia} \tag{10}$$

$$\text{s.t.} \quad \sum_{i,a} (\delta_{ij} - p_{iaj}) x^{(0)}_{ia} = 0,\ \sum_{i,a} x^{(0)}_{ia} = 1,\ x^{(0)}_{ia} \geq 0,$$

$$i \in S, \quad a \in A_0(i).$$

Suppose that $x_*^{(0)}$ is a basic optimal solution for the dual program (10), then any deterministic policy f, such that

$$f_*(i) = a_i, \quad \text{if} \quad x_{*ia_i}^{(0)} > 0,$$

is an average optimal policy for the truncated model. The next theorem states that f_* is also a bias optimal policy for the original MDP.

Theorem 3.1 *Let f_* be any average optimal policy for the truncated Markov decision process. Then,*

$$h(f_*) = h_{opt},$$

where h_{opt} is the optimal value of the bias vector. Namely, the policy f_* is bias optimal.

PROOF: See [7].

4. THE N-DISCOUNT OPTIMALITY

In this section we propose an algorithm for determining n-discount optimal policy based on the system of nested linear programs. This is a generalization of the approach discussed in the preceding section. We remind that n-discount optimality ($n \geq 1$) is more selective criterion which implies average optimal as well as bias optimal criteria.

We will present an iterative approach for calculating the n-discount optimal policies ($n \geq 1$). Namely, the linear program for $(n+1)$-discount optimal policy will be constructed from the solution of the linear program for n-discount optimal policy.

Let us first formally introduce the pair of dual linear programs for the computation of n-discount optimal policy. The primal program can be postulated as follows:

$$\min \ \tilde{g}^{(n)} \tag{11}$$

$$\text{s.t.} \quad \tilde{g}^{(n)} + \sum_j (\delta_{ij} - p_{iaj})\tilde{h}_j^{(n)} \geq -h_{*i}^{(n-1)},$$

$$i \in S, \quad a \in A_n(i).$$

The corresponding dual program is

$$\max \sum_i (-h_{*i}^{(n-1)}) \sum_{a \in A_n(i)} x_{ia}^{(n)} \tag{12}$$

$$\text{s.t.} \quad \sum_{i,a} (\delta_{ij} - p_{iaj}) x_{ia}^{(n)} = 0, \ \sum_{i,a} x_{ia}^{(n)} = 1, \ x_{ia}^{(n)} \geq 0,$$

$$i \in S, \quad a \in A_n(i).$$

Note that the pair of the dual linear programs (9) and (10) for the bias optimality is the base for our induction. Indeed in the case of bias optimality we have $n = 0$, $h_*^{(-1)} := h_*$ and the action set $A_0(i)$ is defined as in (8). From a basic optimal solution of (10) we can construct the bias (or in other words 0-discount) optimal policy.

Now let us perform an induction step. Namely, we are going to construct a linear program for the calculation of $(n+1)$-discount optimal policy.

Firstly define a new truncated set of actions. Let $(g_*^{(n)}, h_*^{(n)})$ be an optimal solution to the problem (11), then $A_{n+1}(i) \stackrel{def}{=}$

$$\{a \in A_n(i) \mid g_*^{(n)} + \sum_j (\delta_{ij} - p_{iaj}) h_{*j}^{(n)} = -h_{*i}^{(n-1)}\}. \tag{13}$$

The induction step is based on the following lemma, which is a generalization of Proposition 3.1.

Lemma 4.1 *Let f be any deterministic and stationary n-discount optimal policy. Then,*
(i) $f(i) \in A_{n+1}(i)$, $i \in S$, and
(ii) the corresponding $(n+1)$-discount vector $y_{n+1}(f)$ (see Laurent expansion (1)) can be calculated by the formula

$$y_{n+1}(f) = [I - P^*(f)] h_*^{(n)} \tag{14}$$

PROOF: We first prove part (i). Let f be a deterministic n-discount optimal policy. Then, according to the induction assumption, it is average optimal for the truncated MDP model associated with linear programs (11) and (12). From Section 2 we know that there exists a one-to-one correspondence between the deterministic average optimal policies and the basic optimal solutions of the dual linear program (12). Hence there exists a basic optimal solution $x_*^{(n)}$ such that $x_{*if(i)}^{(n)} > 0$. Then from the complementary slackness theorem we conclude that

$$g_*^{(n)} + \sum_j (\delta_{ij} - p_{if(i)j}) h_{*j}^{(n)} = -h_{*i}^{(n-1)},$$

that is $f(i) \in A_{n+1}(i)$.

As a result of part (i), we have the following equation valid for any n-discount optimal policy

$$g_*^{(n)} + (I - P(f)) h_*^{(n)} = -h_*^{(n-1)}. \tag{15}$$

We prove the formula (14) by induction. Suppose that for $n = k-1$ it is true. The part (ii) of Proposition 3.1 provides the induction base. Let us prove that the formula (14) holds for $n = k$ as well. Taking into account the structure of coefficients $y_k(f)$ of the Laurent series (1), we have

$$y_{k+1}(f) = (-1)^{k+1} H^{k+2}(f) r(f)$$

$$= -H(f)(-1)^k H^{k+1}(f) r(f) = -H(f) y_k(f).$$

Then, using the induction assumption $y_k(f) = [I - P^*(f)] h_*^{(k-1)}$, formula (15) and the properties of the deviation matrix (3),(4), we obtain

$$y_{k+1}(f) = -H(f) y_k(f) = -H(f)[I - P^*(f)] h_*^{(k-1)}$$

$$= H(f)(-h_*^{(k-1)}) = H(f)[g_*^{(k)} + (I - P(f))h_*^{(k)}]$$

$$= H(f)(I - P(f))h_*^{(k)} = (I - P^*(f))h_*^{(k)}.$$

Thus, according to the principle of the mathematical induction, the formula (14) holds for any $n \geq -1$.

□

It follows from the Definition 1.3 that the policy f is $(n+1)$-discount optimal if it is n-discount optimal and in addition the $(n+1)$-discount vector $y_{n+1}(f)$ achieves its maximal value. Now we can apply the considerations which are similar to those of Section 3. The Lemma 4.1 part (ii) says that the maximization of the vector $y_{n+1}(f)$ is equivalent to the maximization of the product $-P^*(f)h_*^{(n)}$.

The vector $g^{(n+1)}(f) := P^*(f)(-h_*^{(n)})$ can be considered as the expected average reward vector for the Markov decision process with transition matrix $P(f)$, the immediate reward vector $r_{ia}^{(n+1)} := -h_{*i}^{(n)}$ and the truncated action set $A_{n+1}(i)$. Note that $A_{n+1}(i) \neq \emptyset$ for $\forall i \in S$ due to the part (i) of Lemma 4.1. We shall refer to this truncated Markov decision process as $(n+1)$-step truncated MDP model. We emphasize that under Assumption 1.1 the n-step truncated MDP model is completely ergodic for any $n \geq -1$. The optimal value of the expected average reward $g^{(n+1)}(f)$ can be obtained from the next linear program

$$\min \ \tilde{g}^{(n+1)} \tag{16}$$

$$\text{s.t.} \qquad \tilde{g}^{(n+1)} + \sum_j (\delta_{ij} - p_{iaj})\tilde{h}_j^{(n+1)} \geq -h_{*i}^{(n)},$$

$$i \in S, \quad a \in A_{n+1}(i).$$

And the average optimal policy for the $(n+1)$-step truncated model can be determined from the corresponding dual linear program

$$\max \sum_i (-h_{*i}^{(n)}) \sum_{a \in A_{n+1}(i)} x_{ia}^{(n+1)} \tag{17}$$

$$\text{s.t.} \quad \sum_{i,a} (\delta_{ij} - p_{iaj})x_{ia}^{(n+1)} = 0, \quad \sum_{i,a} x_{ia}^{(n+1)} = 1,$$

$$x_{ia}^{(n+1)} \geq 0, \quad i \in S, \quad a \in A_{n+1}(i).$$

Namely, if $x_*^{(n+1)}$ is a basic optimal solution of the dual program (17), then the deterministic policy f_* such that

$$f_*(i) = a_i, \quad \text{if} \quad x_{*ia_i}^{(n+1)} > 0, \tag{18}$$

is average optimal for the $(n+1)$-step truncated model. The next theorem shows that it is also $(n+1)$-optimal for the original problem.

Theorem 4.1 *Let f_* be any average optimal policy for the $(n+1)$-step truncated Markov decision process. Then,*

$$y_{n+1}(f_*) = y_{n+1}^{opt},$$

where y_{n+1}^{opt} is the optimal value of the $(n+1)$-discount vector. Namely, the policy f_* is $(n+1)$-discount optimal.

PROOF: Let d be any stationary $(n+1)$-discount optimal policy for the original Markov decision problem, such a policy exists, see for example [11, 12]. Of course, d is also an n-discount optimal policy. From Lemma 3.1 part (ii), we have

$$y_{n+1}^{opt} = y_{n+1}(d) = h_*^{(n)} - P^*(d)h_*^{(n)}. \tag{19}$$

Since f_* is an average optimal policy for the $(n+1)$-step truncated model, we can write

$$y_{n+1}^{opt} \geq y_{n+1}(f_*) = h_*^{(n)} - P^*(f_*)h_*^{(n)} \geq h_*^{(n)} - P^*(d)h_*^{(n)}. \tag{20}$$

Of course, here $a \geq b$ means $a_i \geq b_i$, for $\forall i \in S$. Then, (19) and (20) imply that

$$y_{n+1}^{opt} = y_{n+1}(f_*),$$

which completes the proof.

□

Thus the system of nested linear programs (11),(12), $n = 0, 1, \ldots$ provides an elegant iterative method for determining n-discount optimal policies. Moreover, this approach eventually leads to the computation of the Blackwell optimal policy.

5. THE ALGORITHM FOR COMPUTATION OF BLACKWELL OPTIMAL POLICY

Here we present a formal algorithm for the computation of the Blackwell optimal policy. Note that n-discount optimal policies, $n = 0, 1, \ldots$, are by-products of that algorithm. Moreover if we want to obtain only the n-discount optimal policy for some particular n, then we should perform at most n main iterative steps.

Algorithm:

1. Solve the dual linear program (6) for determining the average optimal policy. If it has a unique basic optimal solution, then stop. The policy obtained via (7) is Blackwell optimal as well as n-discount optimal for any $n \geq -1$. Otherwise, go to the next step.

2. Compute the solution to the primal program (5).

3. Set $n := 0$.
4. Obtain the truncated action set A_n defined by (8) and (13).
5. Solve the dual program (12) for the n-step truncated MDP model. From a basic optimal solution of (12) using (18) obtain the n-discount optimal policy. If one needs to satisfy only the n-discount optimality criterion, then stop. Also one should stop, if either the basic optimal solution is unique or $n = |S| - 2$. In this case the Blackwell optimal policy is achieved. (Recall that the Blackwell optimality implies the n-discount optimality for any $n \geq -1$.) Otherwise, proceed to the next step.
6. Solve the primal program (11) for the n-step truncated MDP model.
7. Increment the number n by one and return to Step 4.

Note that the Theorem 1.1 guarantees the finite convergence of the above algorithm. The maximal number of repetitions of the main iterative steps 4-7 does not exceed $|S| - 2$.

6. CONCLUSIONS

In this paper we have proposed the algorithm for the computation of n-discount optimal policies as well as the Blackwell optimal policy via the system of nested linear programs. Our algorithm is a generalization of the approach of [2, 7] for computing the bias optimal policy. Since the number of variables in the sequence of nested linear programs is successively reduced, the proposed algorithm is computationally effective. The algorithm can easily be applied to Markov decision processes with Borel state space. The extension of this approach to multi-chain case, namely removing Asumption 1.1, is another possible direction of future research.

7. REFERENCES

[1] D. Blackwell, "Discrete dynamic programming", *Ann. Math. Stat.* **33**, 719-726, 1962.

[2] E. V. Denardo, "Computing a bias-optimal policy in a discrete-time Markov decision problem", *Operations Research*, **18**, pp. 279-289, 1970.

[3] C. Derman, *Finite state Markovian decision processes*, Academic Press, New York, 1970.

[4] O. Hernandez-Lerma and J. B. Lasserre, *Discrete-time Markov control processes. Basic optimality criteria*, Springer-Verlag, 1996.

[5] A. Hordijk, R. Dekker, and L. C. M. Kallenberg, "Sensitivity-analysis in discounted Markovian decision problems", *OR Spektrum*, **7**, pp. 143-151, 1985.

[6] R. G. Jeroslow, "Asymptotic Linear Programming", *Oper. Res.*, v. 21, pp. 1128-1141, 1973.

[7] L. C. M. Kallenberg (1983), *Linear programming and finite Markovian control problems*, Mathematical Centre Tracts 148, Amsterdam.

[8] L. C. M. Kallenberg (1994), "Survey of linear programming for standard and nonstandard Markovian control problems, Part I: Theory", *ZOR - Methods and Models in Operations Research*, **40**, pp. 1-42.

[9] B. F. Lamond and M. L. Puterman, "Generalized inverses in discrete time Markov decision processes", *SIAM J. Matrix Anal. Appl.*, **10**, pp. 118-134, 1989.

[10] B. L. Miller and A. F. Veinott, Jr., "Discrete dynamic programming with a small interest rate", *Ann. Math. Stat.* **40**, pp. 366-370, 1969.

[11] M. L. Puterman (1994), *Markov decision processes*, John Wiley & Sons, New York.

[12] A. F. Veinott, Jr., "Discrete dynamic programming with sensitive discount optimality criteria", *Ann. Math. Stat.* **40**, pp. 1635-1660, 1969.

[13] A. F. Veinott, Jr., "Markov decision chains", in *Studies in Optimization*, eds. G. B. Dantzig and B. C. Eaves, pp. 124-159, 1974.

[14] A. A. Yushkevich, "Blackwell optimal policies in a Markov decision process with a Borel state space", *Z. Oper. Res.*, v. 40, pp. 253-288, 1994.