Asymptotic Linear Programming and Policy Improvement for Singularly Perturbed Markov Decision Processes

Eitan Altman [*] Konstantin E. Avrachenkov [†] Jerzy A. Filar [‡]

Abstract

In this paper we consider a singularly perturbed Markov decision process with finitely many states and actions and the limiting expected average reward criterion. We make no assumptions about the underlying ergodic structure. We present algorithms for the computation of a uniformly optimal deterministic control, that is, a control which is optimal for all values of the perturbation parameter that are sufficiently small. Our algorithms are based on Jeroslow's Asymptotic Linear Programming.

Key Words: Markov Decision Processes, Singular Perturbations, Asymptotic Linear Programming.

Running title: Asymptotic Linear Programming for Singularly Perturbed MDP's.

1 Introduction

Singularly perturbed Markov decision processes (MDPs, for short), are dynamic stochastic systems controlled by a "controller" or a "decision-maker"

[*]INRIA, Centre Sophia-Antipolis, 2004 Route des Lucioles, B. P. 93, 06902 Sophia-Antipolis Cedex, France, e-mail:Eitan.Altman@sophia.inria.fr; Part of this work was done while the author was visiting the University of South Australia and was supported by the Australian Research Council under Grant A49532260

[†]Centre for Industrial and Applicable Mathematics, School of Mathematics, University of South Australia, The Levels, SA 5095, Australia, e-mail: matkea@zarniwoop.levels.unisa.edu.au

[‡]Centre for Industrial and Applicable Mathematics, School of Mathematics, University of South Australia, The Levels, SA 5095, Australia

in which the probability transition law is subject to "small" perturbations that affect the ergodic structure of the underlying Markov chains. A simpler type of perturbation (usually called regular) is one that does not change the ergodic structure.

Both regular and singular perturbations of Markov chains and MDPs have been studied extensively since the 60's (e.g.,[1, 2, 5, 6, 12, 14, 17, 18]). In Abbad and Filar [1] it was demonstrated that singularly perturbed MDPs possess *uniformly optimal* stationary deterministic policies, namely, those which are optimal for all sufficiently small values of the perturbation parameter. However, no practical algorithm for finding such policies was given in [1]. The latter problem is solved in this paper.

It is important to note that the linear programming problem that solves our perturbed MDP has a coefficient matrix that depends on ϵ, the perturbation parameter. Furthermore, at $\epsilon = 0$, there will typically be a change of rank in this coefficient matrix (see Example 5.2). It will be seen that by adapting the *asymptotic linear programming* technique of Jeroslow [10] or by applying similar ideas to the policy iteration method, such difficulties can be overcome and uniformly optimal policies can be found in finitely many steps.

It should be mentioned that asymptotic linear programming has been used in the MDP context by Hordijk et al [7], for the purpose of computing a Blackwell optimal policy in discounted MDP's. Hence it is natural to expect that similar approach might also work in the problem addressed here.

However, there are three important differences between the problem considered in [7] and our model. Firstly, the variable perturbation parameter considered here alters the transition probabilities of the MDP, whereas in [7] the underlying structure is unchanged and only the discount factor is variable. The latter does not affect the ergodic structure and impacts only the overall reward criterion. Secondly, the linear program considered in [7] is simpler than the one analysed in this paper, because it comes from the discounted MDP rather than from the multi-chain limiting average MDP, which is the basis of our problem. This means that we need to consider two sets of constraints and variables. Thirdly, the asymptotics of the problem in [7] (as discount factor tends to one) constitute an important classical sensitivity analysis problem studied by many authors including Blackwell [4], Miller and Veinott [13], Ross and Varadarajan [16], Altman, Hordijk and Kallenberg [3]. In our case, the asymptotics of the perturbed problem (as perturbation tends to zero) have not been fully analysed in the past.

Finally, it has been noted in Huang [9] that the asymptotic linear programming method based on expansions of regular matrix pencils should be applicable to our problem.

2 The model

Consider a perturbed Markov decision process with a finite state space $\mathbf{X} = \{0, 1, ..., N\}$ and a finite *action space* $\mathbf{A}$. Let $\mathbf{A}_x$ denote the set of actions available in state x and $\mathbf{K} = \{(x, a) : x \in \mathbf{A}_x, a \in \mathbf{X}\}$. The probability to go from state x to state y given that action a is used, is given by the transition probability

$$p^{\epsilon}_{xay} := p_{xay} + \epsilon q_{xay} \geq 0 \tag{1}$$

where $\sum_y q_{xay} = 0$ and $\epsilon > 0$ is a "small" perturbation parameter.

A policy u in the *policy space* U is described as $u = \{u_1, u_2, ...\}$, where the decision rule u_t, applied at time epoch t, is a probability measure over $\mathbf{A}$ conditioned on the whole history of actions and states prior to t, as well as on the state at time t. Given an initial distribution β on $\mathbf{X}$, each policy u induces a probability measure denoted by P^u_β on the space of sample paths of states and actions (which serves as the canonical sample space Ω). The corresponding expectation operator is denoted by E^u_β. On this probability space the state and action processes, X_t, A_t, $t = 1, 2, ...$ are defined.

A *Markov policy* $u \in U(M)$ is characterized by the dependence of u_t on the current state x_t and the time t only. A *stationary policy* $g \in U(S)$ is characterized by a single conditional probability measure $p^g_{\cdot|x}$ over $\mathbf{A}$, so that $p^g_{\mathbf{A}|x} = 1$; under g, the state process X_t becomes a Markov chain with transition probabilities given by $P^{\epsilon}_{xy}(g) = \sum_{a \in \mathbf{A}} p^g_{a|x} P^{\epsilon}_{xay}$. The steady state distribution for $\epsilon > 0$ is given by the Cesaro limit

$$P^{*,\epsilon}(g) = \lim_{t\to\infty} \frac{1}{t+1} \sum_{s=0}^{t} [P^{\epsilon}(g)]^s.$$

The class of *deterministic policies* $U(D)$ is a subclass of $U(S)$, and every $g \in U(D)$ is characterized by a mapping $g : \mathbf{X} \to \mathbf{A}$, such that $p^g_{\cdot|x} = \delta_{g(x)}(\cdot)$ is concentrated at the point $g(x)$ for each x.

Let $r : \mathbf{X} \times \mathbf{A} \to \mathbb{R}$, be a (real valued) reward function and define the long-run expected average reward associated with a policy u and with an initial distribution β on $\mathbf{X}$ as

$$R^{\epsilon}_{\beta}(u) = \varliminf_{t\to\infty} \frac{1}{t} E^u_\beta \left[\sum_{s=1}^{t} r(X_s, A_s) \right] \tag{2}$$

Let $OP(\epsilon)$ denote the problem of finding a policy u that maximizes $R^\epsilon_\beta(u)$ for a given initial distribution β. Let R^ϵ_β be the optimal value of $OP(\epsilon)$. A policy that achieves $R^\epsilon_\beta(u) = R^\epsilon_\beta$ is said to be optimal for $OP(\epsilon)$. Let U^ϵ_β denote the set of all such policies.

Following Abbad and Filar [1] we define the limit control problem as the problem of maximizing over $u \in U$

$$R^0_\beta(u) = \underline{\lim}_{\epsilon \to 0} R^\epsilon_\beta(u).$$

However, by Proposition 3.1 in [1] we can restrict consideration to the class of deterministic policies $U(D)$.

A policy u is said to be *limit control optimal*, if for any policy $v \in U(D)$,

$$\underline{\lim}_{\epsilon \to 0}(R^\epsilon_\beta(u) - R^\epsilon_\beta(v)) \geq 0.$$

A policy u is said to be *uniformly optimal*, if for all ϵ sufficiently small and any policy $v \in U(D)$,

$$R^\epsilon_\beta(u) \geq R^\epsilon_\beta(v). \tag{3}$$

The existence of a uniformly optimal stationary deterministic policy was established in Abbad and Filar [1]. A uniformly optimal policy is limit control optimal, but the converse need not hold, as the following example illustrates.

Example 2.1 Consider $\mathbf{X} = \{x, y\}$, $\mathbf{A}_x = \{a, b\}$, $\mathbf{A}_y = \{a\}$; let

$$\begin{array}{ll} p^\epsilon_{xax} = 1, & r(x,a) = 10 \\ p^\epsilon_{xbx} = 1 - \epsilon, & r(x,b) = 10 \\ p^\epsilon_{xby} = \epsilon, & \\ p^\epsilon_{yax} = 1, & r(y,a) = 0 \end{array}$$

Then the stationary policy $u_x = a, u_y = a$ is uniformly optimal with expected average reward $R^\epsilon_\beta(u) = R^\epsilon_\beta = 10$. The stationary policy $v_x = b, v_y = a$ is limit control optimal as $\lim_{\epsilon \to 0} R^\epsilon_\beta(v) = 10$, but for every $\epsilon > 0$,

$$R^\epsilon_\beta(v) = \frac{10}{1+\epsilon} < R^\epsilon_\beta(u).$$

The main purpose of this note is to present two algorithms which derive a uniformly optimal policy.

The following notation is used below: $1\{A\}$ is the indicator function of the set A and $\delta_a(x)$ is the Kronecker delta function. We denote by $|B|$ the cardinality of a set B, i.e. the number of elements in B. For two vectors c and d of appropriate dimensions, $c \cdot d$ denotes the summation over common indices (scalar product). For a stationary policy u, we denote the immediate expected reward vector by $r(u)$, the vector of dimension $|\mathbf{X}|$, whose x-th entry is $r(x,u) := \sum_a r(x,a) u_{a|x}$.

3 Perturbed linear program

One method of solving $OP(\epsilon)$ for a fixed ϵ is based on the solution of a linear program which we present below (see [8]).

Given the initial distribution β over $\mathbf{X}$, define Π_β^ϵ to be the set of $\{(z,\zeta)\}$, $z, \zeta \in \mathbb{R}^{|\mathbf{K}|}$, that satisfy

$$\sum_{y \in \mathbf{X}} \sum_{a \in \mathbf{A}_y} (\delta_y(v) - p_{yav} - \epsilon q_{yav}) z(y,a) = 0, \qquad \forall v \in \mathbf{X} \tag{4}$$

$$\sum_{a \in \mathbf{A}_v} z(v,a) + \sum_{y \in \mathbf{X}} \sum_{a \in \mathbf{A}_y} (\delta_y(v) - p_{yav} - \epsilon q_{yav}) \zeta(y,a) = \beta(v), \qquad \forall v \in \mathbf{X} \tag{5}$$

$$z \geq 0, \qquad \zeta \geq 0. \tag{6}$$

Remark 3.1 *(i) Every $z(\cdot,\cdot) \in \Pi_\beta^\epsilon$ satisfies $\sum_{y,a} z(y,a) = 1$. This can be seen by summing equation (5) over all $v \in \mathbf{X}$.*
(ii) We may delete one of the constraints among (4). This follows from the fact that the coefficients of $z(y,a)$ in (4) sum to 0.

Consider the perturbed linear program: **LP(ϵ):** Find $z, \zeta \in \mathbb{R}^{|\mathbf{K}|}$, that

$$\max\{r \cdot z\}$$

subject to

$$(z,\zeta) \in \Pi_\beta^\epsilon.$$

LP(ϵ) is related to $OP(\epsilon)$ in the following way. Given any $(z, \zeta) \in \Pi^{\epsilon}_{\beta}$, define the stationary policy $g(z, \zeta)$ by

$$p^{g(z,\zeta)}_{a|y} = \begin{cases} \dfrac{z(y,a)}{\sum_{a'} z(y,a')}, & if \ \sum_{a'} z(y,a') > 0 \\ \dfrac{\zeta(y,a)}{\sum_{a'} \zeta(y,a')}, & if \ \sum_{a'} z(y,a') = 0 \ and \ \sum_{a'} \zeta(y,a') > 0 \\ arbitrary, & otherwise. \end{cases} \tag{7}$$

This construction was introduced by Hordijk and Kallenberg [8]. The following lemma is an immediate corollary of the results in [8], for $\epsilon > 0$ and fixed.

Lemma 3.2 *(i) Fix $\epsilon > 0$. The optimal values of $OP(\epsilon)$ and of $LP(\epsilon)$ are equal. (ii) Suppose that $(z^*(\epsilon), \zeta^*(\epsilon))$ is an optimal solution of $LP(\epsilon)$, then $g(z^*(\epsilon), \zeta^*(\epsilon))$ is optimal for $OP(\epsilon)$.*

However, the results in [8] do not permit us to find a uniformly optimal deterministic policy. The latter is a more difficult problem both from a theoretical point of view and due to the fact that the rank of the coefficient matrix of $LP(\epsilon)$ can change at $\epsilon = 0$ (the case of the singular perturbation). This can create numerical problems when $\epsilon > 0$ is small.

4 Asymptotic linear programming

A key step in our method is to find a basis for $LP(\epsilon)$, which is optimal for all ϵ sufficiently small. In order to achieve this we follow an *asymptotic linear programming* approach due to Jeroslow [10]. In particular we consider the entries of the coefficient matrix of $LP(\epsilon)$ not as the elements of the usual Archimedean ordered field of the real numbers, but as elements of the non-Archimedean ordered field $\mathbb{F}(\mathbb{R})$ of rational functions with real coefficients. This approach was previously applied to MDPs by Hordijk et al. [7] to obtain a Blackwell optimal policy (i.e. a policy that is optimal for all discount factor sufficiently close to 1). As in [10, 7], the asymptotic simplex method can also be used to solve our LP(ϵ). An order relation between two rational functions will be defined as in [10]. We denote the ordering and equality in this field by " $>_l$ " and " $=_l$ ", respectively (see Appendix).

The logical steps of asymptotic linear programming method (ALP) are:

- In order to identify an initial feasible basis, we may add the artificial variables $\rho(v), \xi(v), v \in \mathbf{X}$, one variable for each of the constraints in (4) and (5), respectively. We note that

$$(z, \zeta, \rho, \xi) = (\mathbf{0}, \mathbf{0}, \mathbf{0}, \beta)$$

is feasible. A feasible initial basis is an identity matrix corresponding to the artificial variables.

- One may follow the standard or the revised simplex method for solving the linear program $\max\{r \cdot \mathrm{x}\}\ s.t.\ A(\epsilon)\mathrm{x} =_l b, \mathrm{x} \geq 0$, but with respect to the order and equality relations of $\mathbb{F}(\mathbb{R})$. Note that after adding the artificial variables, $A(\epsilon)$ is of full rank. A basis is optimal if and only if the reduced costs $\overline{r}_N(\epsilon) := r_N - r_B B^{-1}(\epsilon) N(\epsilon) \leq_l 0$. Here, r_B (r_N, resp.) is the vector of rewards corresponding to the elements in the basis (not in the basis, resp.); $B(\epsilon)$ is a basic submatrix of $A(\epsilon)$, and $N(\epsilon)$ is the corresponding nonbasic submatrix. Let us use the simplex tableaux. We shall write in the simplex tableaux only the numerators of the rational functions which are the coefficients of variables in $LP(\epsilon)$. This is possible because the entries of the simplex tableaux can be assumed to have common denominators. Let y_{ik} denote the ik-th entry of the current tableau. In the first column we indicate the basic variables. The reduced costs are written in the bottom row of the tableau, and $B^{-1}b$ appears in the right most column. We denote the i-th element of the latter by y_{i0}.

- The entry and exit rules are natural extensions of the usual simplex rules:
  (i) First, we determine the entering non-basic column. Any column (not in the basis) whose corresponding reduced cost is strictly positive (with respect to "$>_l$") can be chosen to enter. As usual, we can choose the column with the largest nonnegative reduced cost to enter.
  (ii) The column p to leave the basis is one that minimizes (with respect to "$>_l$") the ratio $y_{i0}(\epsilon)/y_{ik}(\epsilon)$ among those $i's$ satisfying $y_{ik}(\epsilon) >_l 0$ (k is the index of the entering column).
  We now perform the pivot operation and construct the new tableau. This employs algebraic operations over the field of polynomials $P(\mathbb{R})$.

- The order of the polynomials appearing in both the numerators and denominators never exceeds the number of constraints, that is, twice the number of states. This follows from the fact that (i) each coefficient of the constraints $A(\epsilon)$ is linear in ϵ, and (ii) the elements in the simplex tableau are given by $B^{-1}(\epsilon)b$, where $A(\epsilon)$ is the matrix of constraints, and b is the vector of the right hand side coefficients.

5 Numerical examples

Consider the following illustrative examples.

Example 5.1 Consider $\mathbf{X} = \{x, y\}$, $\mathbf{A}_x = \mathbf{A}_y = \{a, b\}$; let

$$\begin{array}{ll} p^{\epsilon}_{xax} = 1, & r(x,a) = 10 \\ p^{\epsilon}_{xbx} = 1 - \epsilon, & r(x,b) = 10 \\ p^{\epsilon}_{xby} = \epsilon, & \\ p^{\epsilon}_{yax} = 1, & r(y,a) = 0 \\ p^{\epsilon}_{ybx} = p^{\epsilon}_{yby} = 0.5, & r(y,b) = 5. \end{array}$$

We take $\beta(x) = \beta(y) = 0.5$.

By adding the artificial variables, the linear program LP(ϵ) becomes:

$$Max\,\{10z(x,a) + 10z(x,b) + 5z(y,b) - 100\xi(x) - 100\xi(y) - 100\rho(x) - 100\rho(y)\}$$

subject to the constraints

$$\begin{array}{llllll} +\epsilon z(x,b) - z(y,a) - 0.5z(y,b) & & +\xi(x) & & = 0 \\ -\epsilon z(x,b) + z(y,a) + 0.5z(y,b) & & +\xi(y) & & = 0 \\ +z(x,a) + z(x,b) & +\epsilon\zeta(x,b) - \zeta(y,a) - 0.5\zeta(y,b) & & +\rho(x) & = 0.5 \\ \quad +z(y,a) + z(y,b) & -\epsilon\zeta(x,b) + \zeta(y,a) + 0.5\zeta(y,b) & & +\rho(y) & = 0.5 \end{array}$$

$$z(x,a), z(x,b), z(y,a), z(y,b), \zeta(x,a), \zeta(x,b), \zeta(y,a), \zeta(y,b) \geq 0.$$

A reader familiar with MDPs will note that this example is of the so-called "unichain" model (for $\epsilon > 0$). Consequently, a simpler version of LP(ϵ) could have been used (e.g., see Kallenberg [11] p.132). However, the present version of LP(ϵ) applies generally and hence is better suited for demonstrating this new technique.

We added a penalty term for the artificial variables to ensure that they exit the basis. We shall delete the 2nd constraint (and will thus not use $\xi(y)$, see Remark 3.1 (ii)).

The first simplex tableau is given in Table 1. The common denominator is 1. We then choose the 1st column $z(x,a)$ to enter. The row/variable

	artif. variables			*z variables*				*ζ variables*				
Basis	$\xi(x)$	$\rho(x)$	$\rho(y)$	xa	xb	ya	yb	xa	xb	ya	yb	r.h.s.
$\xi(x)$	1	0	0	0	ϵ	-1	-0.5	0	0	0	0	0
$\rho(x)$	0	1	0	$\underline{1}$	1	0	0	0	ϵ	-1	-0.5	0.5
$\rho(y)$	0	0	1	0	0	1	1	0	$-\epsilon$	1	0.5	0.5
red. cost	0	0	0	110	$110 + 100\epsilon$	0	55	0	0	0	0	

Table 1: First tableau of Ex. 5.1

	artif. variables			*z variables*				*ζ variables*				
Basis	$\xi(x)$	$\rho(x)$	$\rho(y)$	xa	xb	ya	yb	xa	xb	ya	yb	r.h.s.
$\xi(x)$	1	0	0	0	ϵ	-1	-0.5	0	0	0	0	0
$z(x,a)$	0	1	0	1	1	0	0	0	ϵ	-1	-0.5	0.5
$\rho(y)$	0	0	1	0	0	1	1	0	$-\epsilon$	$\underline{1}$	0.5	0.5
red. cost	0	-110	0	0	100ϵ	0	55	0	-110ϵ	110	55	

Table 2: Second tableau of Ex. 5.1

to exit is the 2nd one, $\rho(x)$. In all the tableaux the pivoting element is underlined.

The second simplex tableau is given in Table 2. The common denominator is again one. The column that enters the basis is $\zeta(x,a)$ for which the reduced cost 110 is the largest. The column to exit is $\rho(y)$.

The third and fourth simplex tableaux are given in Tables 3, and 4. The common denominator in Table 3 is 1, and in 4 it is ϵ.

At this stage we have obtained an optimal solution over the field of rational functions $\mathbb{F}(\mathbb{R})$ with real coefficients (see Appendix). A uniformly optimal policy uses action a in states x and y, as follows from (7). Note that it is uniformly optimal for all ϵ (satisfying the constraint (1)), i.e. $\epsilon \in (0,1]$. The value of this MDP is 10 (independently of ϵ). The stationary deterministic policies that choose action b in state x are optimal for the limit problem but are not optimal for any positive ϵ.

Now consider the second example, which exposes the singularity of the problem.

Basis	artif. variables			z variables				ζ variables				
	$\xi(x)$	$\rho(x)$	$\rho(y)$	xa	xb	ya	yb	xa	xb	ya	yb	r.h.s.
$\xi(x)$	1	0	0	0	$\underline{\epsilon}$	-1	-0.5	0	0	0	0	0
$z(x,a)$	0	1	1	1	1	1	1	0	0	0	0	1
$\zeta(y,a)$	0	0	1	0	0	1	1	0	$-\epsilon$	1	0.5	0.5
red. cost	0	-110	-110	0	100ϵ	-110	-55	0	0	0	0	

Table 3: Third tableau of Ex. 5.1

Basis	artif. variables			z variables				ζ variables				
	$\xi(x)$	$\rho(x)$	$\rho(y)$	xa	xb	ya	yb	xa	xb	ya	yb	r.h.s.
$z(x,b)$	1	0	0	0	ϵ	-1	-0.5	0	0	0	0	0
$z(x,a)$	-1	ϵ	ϵ	ϵ	0	$1+\epsilon$	$0.5+\epsilon$	0	0	0	0	ϵ
$\zeta(y,a)$	0	0	ϵ	0	0	ϵ	ϵ	0	$-\epsilon^2$	ϵ	0.5ϵ	0.5ϵ
red. cost	-100ϵ	$-110\ \epsilon$	-110ϵ	0	0	-10ϵ	-5ϵ	0	0	0	0	

Table 4: Fourth tableau of Ex. 5.1

Example 5.2 *Consider $X = \{w, x, y\}$ and $A_w = \{a, b\}$, $A_x = \{a\}$, $A_y = \{a\}$. And let*

$$\begin{array}{ll} p^\epsilon_{waw} = 1-\epsilon, p^\epsilon_{way} = \epsilon, & r(w,a) = 10 \\ p^\epsilon_{wbw} = 1-\epsilon, p^\epsilon_{wbx} = \epsilon, & r(w,b) = 10 \\ p^\epsilon_{xax} = 1-\epsilon, p^\epsilon_{xay} = \epsilon, & r(x,a) = 10 \\ p^\epsilon_{yaw} = p^\epsilon_{yax} = 0.5, & r(y,a) = 0. \end{array}$$

Since the perturbed Markov chain is ergodic for all stationary policies, we may use a simplified form of LP(ϵ) with only the z-variables (see Kallenberg [11], p.128). Thus our LP(ϵ) problem is

$$\max\{\sum_{x,a} r(x,a)z(x,a) | \mathcal{A}(\epsilon)z = b, z \geq 0\},$$

where

$$\mathcal{A}(\epsilon) = \begin{pmatrix} \epsilon & \epsilon & 0 & -0.5 \\ 0 & -\epsilon & \epsilon & -0.5 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

	z variables				
Basis	$(1,a)$	$(1,b)$	$(2,a)$	$(3,a)$	r.h.s.
$z(1,a)$	$1+\epsilon$	$\underline{1.5+\epsilon}$	0	0	0.5
$z(2,a)$	0	$-0.5-\epsilon$	$1+\epsilon$	0	0.5
$z(3,a)$	0	ϵ	0	$1+\epsilon$	ϵ
red. cost	0	$10\ \epsilon$	0	0	

Table 5: First tableau of Ex. 5.2

We shall start with a basis consisting of the columns that correspond to the policy (a,a,a). Hence

$$B(\epsilon) = \begin{pmatrix} \epsilon & 0 & -0.5 \\ 0 & \epsilon & -0.5 \\ 1 & 1 & 1 \end{pmatrix}.$$

We have

$$B^{-1}(\epsilon) = (\epsilon+\epsilon^2)^{-1} \begin{pmatrix} 0.5+\epsilon & -0.5 & 0.5\epsilon \\ -0.5 & 0.5+\epsilon & 0.5\epsilon \\ -\epsilon & -\epsilon & \epsilon^2 \end{pmatrix}.$$

Note that $B^{-1}(\epsilon)$ has a singularity at $\epsilon = 0$. Now,

$$B^{-1}(\epsilon)b = (1+\epsilon)^{-1} \begin{pmatrix} 0.5 \\ 0.5 \\ \epsilon \end{pmatrix},$$

and finally

$$B^{-1}(\epsilon)\mathcal{A}_{[1,b]}(\epsilon) = (1+\epsilon)^{-1} \begin{pmatrix} 1.5+\epsilon \\ -0.5-\epsilon \\ \epsilon \end{pmatrix}.$$

The reduced reward coefficients corresponding to the element $(1,b)$ is $r(1,b) - (r^B B^{-1}(\epsilon)\mathcal{A}_{[1,b]}(\epsilon)) = 10\epsilon(1+\epsilon)^{-1}$ (where r^B is the vector of rewards corresponding to the elements in the basis).

In the first simplex tableau (Table 5) we only write the numerators. The denominator which is common to all entries is $(1+\epsilon)$. The expected long-run average reward corresponding to this basis is $10(1+\epsilon)^{-1}$.

To obtain the second tableau, the second column enters the basis and the first row exits the basis. This simply corresponds to the policy (b,a,a).

	z variables				
Basis	$(1,a)$	$(1,b)$	$(2,a)$	$(3,a)$	r.h.s.
$z(1,b)$	$1+\epsilon$	$1.5+\epsilon$	0	0	0.5
$z(2,a)$	$0.5+\epsilon$	0	$1.5+\epsilon$	0	1
$z(3,a)$	$-\epsilon$	0	0	$1.5+\epsilon$	ϵ
red. cost	- 10 ϵ	0	0	0	

Table 6: Second tableau of Ex. 5.2

The common denominator to all entries in the second Tableau is $(1.5+e)$, which we have omitted.

The reward corresponding to the new basis is $15(1.5+\epsilon)^{-1}$ and the policy (b,a,a) is uniformly optimal.

6 Asymptotic policy iteration

The method of Jeroslow [10] can also be used in a policy iteration type method for obtaining a uniformly optimal policy. The steps of *Asymptotic Policy Iteration Method* (API) are as follows:

1. Start with an arbitrary stationary deterministic policy u.

2. Solve the equations with unknowns R, H, W, which are all $|\mathbf{X}|$-dimensional vectors whose entries are elements in the field $\mathbb{F}(\mathbb{R})$ of rational functions (in ϵ) with real coefficients:

$$P^\epsilon(u)R =_l R \tag{8}$$

$$(P^\epsilon(u) - I)H =_l R - r(u) \tag{9}$$

$$(P^\epsilon(u) - I)W =_l H \tag{10}$$

   (R and H are uniquely determined (e.g., see [15] p. 452), and $R =_l R^\epsilon(u), H =_l H^\epsilon(u)$.)

3. Choose a stationary deterministic policy v that selects an action $v(x) = a$ in state x as follows: If there is an action a such that

$$\sum_y p^\epsilon_{xay} R_y >_l R_x,$$

then $v(x) = a$. If such an action does not exist, but there is an action a for which

$$\sum_y p^\epsilon_{xay} R_y =_l R_x \quad \text{and} \quad r(x,a) + \sum_y p^\epsilon_{xay} H_y >_l R_x + H_x,$$

then set $v(x) := a$. If there are no actions satisfying either of the above conditions, then the action at this state is unchanged (we set $v(x) := u(x)$).

4. If $u = v$ then stop. Otherwise, set $u := v$, and return to Step 1.

7 Finite convergence

The next proposition demonstrates that both the (ALP) and (API) methods indeed solve the original problem (3).

Proposition 1 *Consider a singularly perturbed MDP and the associated optimal control problem $OP(\epsilon)$. Apply either the (ALP) or the (API) method to this problem. Both methods terminate in finitely many iterations and yield a uniformly optimal control.*

PROOF: (i) For the (API) method.
In Steps 2 and 3 of (API) the relations " $>_l$ " and " $=_l$ " are with respect to the ordering of rational function with real coefficients, as in [10, 7] (see exact definitions given in the Appendix). The above algorithm finds an optimal policy in finitely many steps. This follows from the fact that each step in the algorithm yields a strict improvement of the long-run expected average reward criterion with respect to the ordering over $\mathbb{F}(\mathbb{R})$. Now, the uniformity follows from the following reasoning. Consider a rational function p/q, where p and q are polynomials with rational coefficients. It is called positive (nonnegative, resp.) if $p/q >_l (\geq_l) p_0$, respectively. Then p/q is positive if and only if there exists $\epsilon_0 > 0$ such that $p(\epsilon)/q(\epsilon) > 0$ for all $\epsilon \in (0, \epsilon_0]$. The latter implies that we find a policy which is optimal for all sufficiently small ϵ, namely, a uniform optimal policy.

(ii) For the (ALP) method.
This follows immediately from the arguments presented by Jeroslow [10].

8 Appendix: The field of rational functions with real coefficients

In order for this paper to be self contained, we include a brief description of the field of rational functions (see e.g. [7]). Let $P(\mathbb{R})$ denote the set of polynomials with real coefficients of the form

$$p(x) = a_0 + a_1 x + ... + a_n x^n, \qquad n \in \mathbb{N}.$$

Let p_0 and p_1 denote the polynomials $p_0 \stackrel{\text{def}}{=} 0, p_1 \stackrel{\text{def}}{=} 1$, respectively. Let $d(p)$ denote the dominating (leading) coefficient of the polynomial p; it is a_k, where k is the smallest integer with $a_k \neq 0$.

The field $\mathbb{F}(\mathbb{R})$ of rational functions with real coefficients consists of elements of the form p/q, where $p, q \in P(\mathbb{R})$. A polynomial p is identified with $p/1$. We say that $p/q =_l r/s$ if $ps = qr$, $p, q, r, s \in P(\mathbb{R})$. The addition and multiplication of elements in $\mathbb{F}(\mathbb{R})$ are defined by

$$\frac{p}{q} + \frac{r}{s} =_l \frac{ps + rq}{qs}, \qquad \frac{p}{q} \cdot \frac{r}{s} =_l \frac{pr}{qs}, \qquad p, q, r, s \in P(\mathbb{R}).$$

p_0 and p_1 are the identities with respect to addition and multiplication, respectively. A complete ordering on $\mathbb{F}(\mathbb{R})$ is obtained by defining

$$p/q >_l p_0 \quad \text{if and only if} \quad d(p)d(q) > 0.$$

References

[1] Abbad M, Filar JA (1992) Perturbation and stability theory for Markov control problems. IEEE Trans. on Automatic Control 37:1415-1420

[2] Abbad M, Filar JA, Bielecki TR (1992) Algorithms for singularly perturbed limiting average Markov control problems. IEEE Trans. on Automatic Control 37:1421-1425

[3] Altman E, Hordijk A, Kallenberg LCM (1996) On the value function in constrained control of Markov chains. Mathematical Methods of Operations Research 44:387-399

[4] Blackwell D (1962) Discrete dynamic programming. Ann. Math. Statist. 33:719-726

[5] Delebecque F, Quadrat JP (1981) Optimal control of Markov chains admitting strong and weak interactions. Automatica 17:281-296

[6] Delebecque F (1983) A reduction process for perturbed Markov chains. SIAM J. App. Math. 48:325-350

[7] Hordijk A, Dekker R, Kallenberg LCM (1985) Sensitivity analysis in discounted Markovian decision problems. Spectrum 7:143-151

[8] Hordijk A, Kallenberg LCM (1979) Linear programming and Markov decision chains. Management Science 25:352-362

[9] Huang Y (1996) A canonical form for pencils of matrices with applications to asymptotic linear programs. Linear Algebra Appl. 234:97-123

[10] Jeroslow RG (1973) Asymptotic linear programming. Oper. Res. 21:1128-1141

[11] Kallenberg LCM (1983) Linear programming and finite markovian control problems. Mathematisch Centrum, Amsterdam

[12] Lasserre JB (1994) A formula for singular perturbation of Markov chains. J. Appl. Prob. 31:829-833

[13] Miller BL, Veinott AF (1969) Discrete dynamic programming with a small interest rate. Ann. Math. Statist. 40:366-370

[14] Pervozvanskii AA, Gaitsgori VG (1988) Theory of suboptimal decisions. Kluwer Academic Publishers, Dordrecht, Netherlands

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

[16] Ross KW, Varadarajan R (1991) Multichain Markov decision processes with a sample path constraint: a decomposition approach. Math. of Oper. Res. 16:195-201

[17] Schweitzer PJ (1968) Perturbation theory and finite Markov chains. J. App. Prob. 5:401-413

[18] Schweitzer PJ (1986) Perturbation series expansions for nearly completely-decomposable Markov chains. In Boxma OJ, Cohen JW, Tijms HC (eds) Teletrafic Analysis and Computer Performance Evaluation. Elsevier Science Publishers B.V. (North-Holland) pp 319-328