An Asymptotic Simplex Method and Markov Decision Processes

Eitan Altman [*], Konstantin Avrachenkov [†], Jerzy Filar [‡]

January 16, 2002

Abstract

The purpose of this paper is twofold. First we present an asymptotic simplex method for the parametric linear programming. The asymptotic simplex method allows to find a solution which is optimal for all sufficiently small values of the parameter. The present implementation of the asymptotic simplex method is based on the Laurent series expansions and on the efficient use of the recurrent formulae for the Laurent series coefficients. In the second part of the paper we apply the asymptotic simplex method to several models of the Markov decision processes: computation of Blackwell optimal policies, Markov branching decision chains, singularly perturbed MDPs and constrained MDPs and stochastic games.

1 Introduction

The present paper consists of two parts. The first part is devoted to the asymptotic simplex method, which is designed to solve the parametric linear program of the following form

$$\min_x \{(c^{(0)} + \varepsilon c^{(1)})x\} \tag{1}$$

subject to

$$(A^{(0)} + \varepsilon A^{(1)})x = b^{(0)} + \varepsilon b^{(1)}, \quad x \geq 0. \tag{2}$$

[*]INRIA Sophia Antipolis, 2004 Route des Lucioles, B.P. 93, 06902, Sophia Antipolis Cedex, France, e-mail: eitan.altman@sophia.inria.fr

[†]INRIA Sophia Antipolis, 2004 Route des Lucioles, B.P. 93, 06902, Sophia Antipolis Cedex, France, e-mail: k.avrachenkov@sophia.inria.fr

[‡]School of Mathematics, University of South Australia, The Levels, SA 5095, Australia, e-mail: jerzy.filar@unisa.edu.au

where $A^{(0)}, A^{(1)} \in \mathbb{R}^{m\times n}$, $b^{(0)}, b^{(1)} \in \mathbb{R}^{m\times 1}$, $c^{(0)}, c^{(1)} \in \mathbb{R}^{1\times n}$ and ε is a small parameter. The parametric linear program (1), (2) is often referred to as *perturbed* linear program. We are especially interested in the case of singular perturbation when an optimal solution of the perturbed program does not converge to an optimal solution of the unperturbed program. This typically hapens when $rank(A^{(0)} + \varepsilon A^{(1)}) > rank(A^{(0)})$ for $\varepsilon > 0$ and sufficiently small. Many problems in the theory of Markov Decision Processes (MDPs) can be formulated in terms of the linear parametric programming. Thus, in the second part of the paper we explore the application of the asymptotic simplex method to MDP problems.

The general idea of the asymptotic simplex method was probably first given by Jeroslow [14, 15]. He proposed a simplex-like method which works directly over the field of rational functions. The main disadvantage of this method is that the polynomials involved in calculation can have a high degree. For instance, the inversion of a basis matrix takes $O(m^4 log(m))$ flops [8]. It can be shown that Jeroslow's method can be viewed as an instance of a more general theory of extending algorithms for parametric problems (see Eaves and Rothblum [6]). Another application of this general theory would be a method which works in the field of power series expansions. The basic idea of the asymptotic simplex method which works over the field of power series was given by Pervozvanskii and Gaitsgory [21]. However they imposed some strong restrictions. For example, it was required that the inverse of the perturbed basis matrix has a pole of order one at the singular point $\varepsilon = 0$. Later Lamond [16] proposed a method for the expansion of the inverse of the basis matrix which demands $O(\bar{s}m^3)$ flops, where $\bar{s}$ is the maximal order of poles of the basis matrices. Huang [12] improved further the expansion of the perturbed basis matrix by proposing an algorithm which demands only $O(m^3)$ flops. In this paper we propose an updating procedure which deals with the general case and demands only $O(\bar{s}m^2)$. It is based on the elegant recurrent formulae of Langenhop [18] and Schweitzer and Stewart [23]. We also use the recurrent formulae of [18, 23] to efficiently calculate Laurent series terms of the simplex method quantities.

2 An Asymptotic Simplex method

In this Section we present the asymptotic simplex method for obtaining an a solution to the parametric linear program (1),(2) which is optimal for all sufficiently small positive ε.

At first let us introduce some rather mild assumptions, which can be relaxed even further [7]. Let M_ε denote the feasible region (2) and M_0 be the feasible region of the unperturbed problem with $\varepsilon = 0$.

Assumption 1 The region M_0 is bounded.

The above assumption ensures that basic feasible solutions of the perturbed program (1),(2) can be expanded as Taylor series (see [7] for details).

Assumption 2 The perturbed problem is non-degenerate, namely, every element of the basic feasible vector $x_B(\varepsilon) = A_B^{-1}(\varepsilon)b(\varepsilon), \varepsilon \in (0, \overline{\varepsilon}]$ is positive.

Recall that in the classical revised simplex method [19] we see that everything depends on the inverse basis matrix A_B^{-1}. In the parametric case it is necessary to realise that the analogous matrix $A_B^{-1}(\varepsilon)$ is a Laurent series whose structure and coefficients determine the asymptotic behaviour, as $\varepsilon \to 0$. In particular,

$$A_B^{-1}(\varepsilon) = \frac{1}{\varepsilon^s} B^{(-s)} + ... + \frac{1}{\varepsilon} B^{(-1)} + B^{(0)} + ... \;. \tag{3}$$

If $A_B(\varepsilon)$ becomes singular at $\varepsilon = 0$, then the above series will have a pole of order s at $\varepsilon = 0$ and will contain a nontrivial *singular part*; defined by

$$A_B^S(\varepsilon) = \frac{1}{\varepsilon^s} B^{(-s)} + ... + \frac{1}{\varepsilon} B^{(-1)} \;. \tag{4}$$

Similarly, a *regular part* of (3) is defined by

$$A_B^R(\varepsilon) = B^{(0)} + \varepsilon B^{(1)} + ... \;. \tag{5}$$

Clearly, if $A_B^S(\varepsilon) \neq 0$ and ε is small then standard simplex operations could result in unstable behaviour. The asymptotic simplex method presented here overcomes this difficulty by working with the coefficients of (3).

At first sight it might appear that computations involving the series expansion (3) would be too difficult. Fortunately, recursive formulae developed by Langenhop [18] and Schweitzer and Stewart [23] provide tools that can be adapted to the revised simplex method. A key observation here is that if $B^{(0)}$ and $B^{(-1)}$ are known, then every coefficient of (3) can be obtained according to

$$B^{(k)} = (-B^{(0)} A_B^{(1)})^{k-1} B^{(0)}, \quad k = 0, 1, ... \tag{6}$$

$$B^{(-k)} = (-B^{(-1)} A_B^{(0)})^{k-1} B^{(-1)}, \quad k = 1, ..., s, \tag{7}$$

where $A_B^{(i)}$ are the coefficients of ε^i in the basis matrix $A_B(\varepsilon)$. Consequently, if $B^{(0)}$ and $B^{(-1)}$ can be efficiently updated, when moving from one basis to another, then any coefficient of (3) can also be calculated for the next basis via (6),(7), if it is needed.

However, we still need to compute $B^{(0)}$ and $B^{(-1)}$ for the initial step of the asymptotic simplex method. There are two solutions to this problem. The first solution is to compute the singular part and the first regular coefficient of asymptotic expansion (3) by using methods presented in [3]. The other solution is to start the asymptotic simplex method with an analog of phase 1 method [4, 19]. Note that when we introduce artificial variables in phase 1, the constraints of the unperturbed program become linearly independent even in the case of singular perturbation. The later enables us to start from a basis matrix which possesses a Taylor series expansion instead of a Laurent series. This significantly facilitates computations. In addition, if we use the phase 1 method, we need not be concerned about the choice of an initial basic feasible solution.

In asymptotic simplex method we need to compare two functions in some small neighbourhood of zero. This will be done with the help of asymptotic expansion (3) and the concept of lexicographical ordering. Before proceeding further, let us recall the definition of *lexicographical ordering* [25, 26].

Definition 1 A vector a (possibly infinite) is called *lexicographically nonnegative*, written $a \succeq 0$, if the first nonzero element (if any) in a is positive, and a is called *lexicographically positive*, written $a \succ 0$, if $a \succeq 0$ and $a \neq 0$. For two vectors a and b we say that a is *lexicographically greater (strictly)* than b, if $a - b \succeq 0$ $(a - b \succ 0)$.

The following lemma [16, 12] is also very useful in our analysis.

Lemma 1 *Suppose $c(\varepsilon) = a(\varepsilon)/b(\varepsilon)$ is a rational function with the degrees of the polynomials $a(\varepsilon)$ and $b(\varepsilon)$ being m and n, respectively. Then the function $c(\varepsilon)$ can be expanded as a Laurent series*

$$c(\varepsilon) = \frac{1}{\varepsilon^s} c^{(-s)} + \frac{1}{\varepsilon^{s-1}} c^{(-s+1)} + ...$$

in some punctured neighbourhood of zero with the order of pole s that is at most n. Moreover, if $c^{(-s)} = c^{(-s+1)} = ... = c^{(m)} = 0$, then $c(\varepsilon) \equiv 0$.

Suppose that we have an analytic function $g(\varepsilon)$ which can be expanded as a Laurent series at $\varepsilon = 0$ with the finite singular part

$$g(\varepsilon) = \frac{1}{\varepsilon^{-s}} g^{(-s)} + ... + \frac{1}{\varepsilon} g^{(-1)} + g^{(0)} + \varepsilon g^{(1)} + ... \ .$$

We construct from the coefficients of the above series the infinite vector

$$\gamma = [g^{(-s)}, ..., g^{(0)}, g^{(1)}, ...].$$

It is easy to see that $g(\varepsilon) > 0$ for ε sufficiently small and positive $\Leftrightarrow \gamma \succ 0$. Moreover, if $g(\varepsilon)$ is a rational function then only a finite number of elements in γ needs to be checked (see Lemma 1). The comparison (in a neighbourhood of 0) between two functions $g(\varepsilon)$ and $h(\varepsilon)$ possessing Laurent series expansions with finite order poles can be carried out similarly, by considering $g(\varepsilon) - h(\varepsilon)$.

Now let us formally present the *asymptotic simplex method* and comment on its possible implementation. The proof of its finite convergence is given in [7].

The asymptotic simplex method:
Let the initial set of basic indices B be given.

Step 1: Obtain or update only the singular and the first regular coefficients of the Laurent series expansion (3) for the inverse basis matrix $A_B^{-1}(\varepsilon)$. The implementation of this step is discussed at the end of this section.

Step 2 (idea): In asymptotic simplex method we have to decide which column enters the basis. Namely, among the non-basic elements of the reduced cost vector

$$r_N(\varepsilon) := \lambda(\varepsilon) A_N(\varepsilon) - c_N(\varepsilon), \tag{8}$$

where $\lambda(\varepsilon) := c_B(\varepsilon) A_B^{-1}(\varepsilon)$, we need to find such k that

$$k \in \arg\max_{j \in N} \{ r_j(\varepsilon) | r_j(\varepsilon) > 0, \varepsilon \in (0, \bar{\varepsilon}] \}.$$

Substituting (3) into (8), we obtain the next asymptotic expansion

$$r_N(\varepsilon) = \frac{1}{\varepsilon^s} r_N^{(-s)} + \frac{1}{\varepsilon^{s-1}} r_N^{(-s+1)} + \dots \quad ,$$

where $r_N^{(i)} := \lambda^{(i)} A_N^{(0)} + \lambda^{(i-1)} A_N^{(1)} - \delta_{0i} c_N^{(0)} - \delta_{1i} c_N^{(1)}$ and $\lambda^{(i)} := c_B^{(0)} B^{(i)} + c_B^{(1)} B^{(i-1)}$. Let us construct the following (semi-infinite) matrix

$$\rho = \begin{bmatrix} r_N^{(-s)} \\ r_N^{(-s+1)} \\ \dots \end{bmatrix}$$

and denote its i-th column by ρ_i. As mentioned above, the lexicographical ordering can be used to compare functions in the "small" neibourhood near zero. In particular, it is easy to see that

$$\arg\max_{j \in N} \{ r_j(\varepsilon) | r_j(\varepsilon) > 0, \varepsilon \in (0, \bar{\varepsilon}] \} =$$

$$= \arg \text{lex-}\max_{j \in N} \{\rho_j | \rho_j \succ 0\},$$

where "lex-max" is a maximum with respect to the lexicographical ordering and "arg lex-max" is an index at which "lex-max" is attained. Note that to compare two reduced cost coefficients $r_i(\varepsilon)$ and $r_j(\varepsilon)$ for sufficiently small ε we need only to check a *finite* number of elements of the vectors ρ_i and ρ_j. This follows from the fact that $r_i(\varepsilon)$ and $r_j(\varepsilon)$ are rational functions (see Lemma 1). A practical implementation of the lexicographical entering rule is as follows:

Step 2 (implementation): Set $i := -s$ and $\lambda^{(-s-1)} := 0$.

Step 2a) Calculate the i-th term of the Laurent expansion for the vector of simplex multipliers

$$\lambda^{(i)} := c_B^{(0)} B^{(i)} + c_B^{(1)} B^{(i-1)}.$$

Note that by using the recurrent formulae (6),(7), one can efficiently compute $\lambda^{(i)}$. For instance, $\lambda^{(i)} = \lambda^{(i-1)}(-A_B^{(1)} B^{(0)})$ for $i \geq 2$.

Step 2b) Calculate the i-th term of the Laurent expansion for the vector of non-basic reduced cost coefficients

$$r_N^{(i)} := \lambda^{(i)} A_N^{(0)} + \lambda^{(i-1)} A_N^{(1)} - \delta_{0i} c_N^{(0)} - \delta_{1i} c_N^{(1)}$$

where δ_{0i} and δ_{1i} are the Kronecker deltas. Let $N^{(-s-1)} := N$ and $N^{(i)} = \{j \in N^{(i-1)} | r_j^{(i)} = 0\}$. If $r_j^{(i)} < 0$ for $j \in N^{(i-1)}$, then STOP; the current solution is optimal for all sufficiently small positive ε. If there is an index k such that

$$k := \arg \max_{j \in N^{(i-1)}} \{r_j^{(i)} | r_j^{(i)} > 0\},$$

then k identifies the entering non-basic variable; go to Step 3.

Step 2c) If the algorithm does not stop and k has not been identified in Step 2b, and $i < m+1$, then $N^{(i)}$ is not empty. Increment the index i and return to Step 2a to consider the higher order approximation of $r_N(\varepsilon)$.

Remark 1 *Lemma 1 ensures that if index $i = m+1$ is reached in Step 2c and $N^{(m+1)}$ is still non-empty, then $r_j(\varepsilon) \equiv 0$ for $j \in N^{(m+1)}$ and ε sufficiently small. The latter implies that $r_N(\varepsilon) \leq 0$ for all ε sufficiently small, that is, the current solution is optimal for all $\varepsilon > 0$ and sufficiently small.*

Step 3 (idea): Now, as in the revised simplex method, we have to find out which elements of the vector $y_k(\varepsilon) = A_B^{-1}(\varepsilon) a_k(\varepsilon)$ are positive for $\varepsilon > 0$ and sufficiently small. Namely, we have to identify the set of indices $P :=$

$\{l | [y_k(\varepsilon)]_l > 0, \varepsilon \in (0, \bar{\varepsilon}]\}$. Towards this, as in Step 2, we first expand $y_k(\varepsilon)$ as a Laurent series

$$y_k(\varepsilon) = \frac{1}{\varepsilon^s} y_k^{(-s)} + \frac{1}{\varepsilon^{s-1}} y_k^{(-s+1)} + ...$$

and then define an auxiliary semi-infinite matrix

$$\upsilon = [y_k^{(-s)}, y_k^{(-s+1)}, ...].$$

Let υ_l denote the l-th row of matrix υ. Then the set P can be described as $P = \{l | \upsilon_l \succ 0\}$ and the practical procedure for its determination is

Step 3 (implementation): Set $i := -s$, $B^{(-s-1)} := 0$ and $P = \emptyset$.

Step 3a) Calculate the i-th term of the Laurent expansion for $y_k(\varepsilon)$

$$y_k^{(i)} = B^{(i)} a_k^{(0)} + B^{(i-1)} a_k^{(1)}.$$

Step 3b) Let $Q^{(-s-1)} := \{1, ..., m\}$ and $Q^{(i)} := \{j \in Q^{(i-1)} | [y_k^{(i)}]_j = 0\}$. Add the index $j \in Q^{(i-1)}$ to the set P if $[y_k^{(i)}]_j > 0$. If $Q^{(i)} = \emptyset$, then go to Step 3d.

Step 3c) If $Q^{(i)} \neq \emptyset$ and $i < m$, then increment i by one and return to Step 3a. However, if the index i reached m, then go to the next Step 3d. Lemma 1 guarantees that $[y_k(\varepsilon)]_j \equiv 0, j \in Q^{(m)}$.

Step 3d) Stop. At this point the set P is determined.

Remark 2 *From Assumption 1 one can deduce [7] that the feasible region for the perturbed problem is bounded. The latter implies that in our setting the set P is always non-empty.*

Step 4 (idea): Now we have to choose a basic variable which exits the basis, namely we have to find

$$p \in \arg\min_l \left\{ \frac{[x_B(\varepsilon)]_l}{[y_k(\varepsilon)]_l} \;\; | \;\; l \in P, \varepsilon \in (0, \bar{\varepsilon}] \right\}.$$

According to the previous step the functions $[y_k(\varepsilon)]_l, l \in P$ can be expressed as a Laurent series

$$[y_k(\varepsilon)]_l = \varepsilon^{q_l} [y_k^{(q_l)}]_l + \varepsilon^{q_l+1} [y_k^{(q_l+1)}]_l + ..., \tag{9}$$

with $y_l^{(q_l)} > 0$. Due to the non-degeneracy Assumption 2 $[x_B(\varepsilon)]_l$ can be expressed as power series with a positive leading coefficient

$$[x_B(\varepsilon)]_l = \varepsilon^{t_l} x_{Bl}^{(t_l)} + \varepsilon^{t_l+1} x_{Bl}^{(t_l+1)} + ..., \quad x_{Bl}^{(t_l)} > 0. \tag{10}$$

Since $[x_B(\varepsilon)]_l$ and $[y_k(\varepsilon)]_l$ are rational functions, the quotient $\Delta_l(\varepsilon) := [x_B(\varepsilon)]_l/[y_k(\varepsilon)]_l$ can also be written in terms of Laurent series

$$\Delta_l(\varepsilon) = \varepsilon^{t_l - q_l}(\Delta_l^{(0)} + \varepsilon\Delta_l^{(1)} + \varepsilon^2\Delta_l^{(2)} + ...), \tag{11}$$

where the coefficients $\Delta_l^{(i)}$ are calculated by simple recurrent formulae (see below Step 4c). Thus, similarly to the previous steps the lexicographical ordering can be implemented as follows:

Step 4 (implementation): Set $i := 0$.

Step 4a) Form the set of indices corresponding to the maximal powers of the leading coefficients in (11).

$$R^{(-1)} := \{j | j = \arg\max_l \{t_l - q_l | l \in P\}\}$$

Step 4b) Calculate the $(q_l + i)$-th and $(t_l + i)$-th terms of the expansions (9),(10) respectively

$$[y_k^{(q_l+i)}]_l = [B^{(q_l+i)}]_l a_k^{(0)} + [B^{(q_l+i+1)}]_l a_k^{(1)}, \quad l \in R^{(i-1)},$$

and

$$x_{Bl}^{(t_l+i)} = [B^{(t_l+i)}]_l b^{(0)} + [B^{(t_l+i+1)}]_l b^{(1)}, \quad l \in R^{(i-1)},$$

Step 4c) Calculate the i-th coefficient of expansion (11)

$$\Delta_l^{(i)} = (x_{Bl}^{(t_l+i)} - \sum_{j=0}^{i-1} [y_k^{(q_l+i-j)}]_l \Delta_l^{(j)})/[y_k^{(q_l)}]_l, \quad l \in R^{(i-1)}.$$

Step 4d) Form the following set of indices

$$R^{(i)} := \{j | j = \arg\min_l \{\Delta_l^{(i)} | l \in R^{(i-1)}\}\}.$$

If $R^{(i)}$ consists of a unique index p, then go to Step 5. If $R^{(i)}$ is not a singleton and $i < 2m+1$, then we should take into account the higher order approximation of $\Delta_l(\varepsilon)$. Namely, increment i by one and return to Step 4b. However, if $i = 2m+1$ then choose any $p \in R^{(i)}$ and go to Step 5.

Remark 3 *Again Lemma 1 guarantees that $\Delta_p(\varepsilon) \equiv \Delta_q(\varepsilon)$, if $p, q \in R^{(2m+1)}$.*

Step 5: Construct a new basis $A_{B'}(\varepsilon)$ obtained from $A_B(\varepsilon)$ by replacing $a_p(\varepsilon)$ with $a_k(\varepsilon)$. Go to Step 1.

This completes the algorithm.

As in the revised simplex method, it is possible to update the expansion (3) for the inverse of the new basis matrix $A_{B'}(\varepsilon)$ via the multiplication of the series by $E(\varepsilon) = [e_1, ..., e_{p-1}, \xi(\varepsilon), e_{p+1}, ..., e_m]$, where $\xi(\varepsilon) = [-\frac{y_1(\varepsilon)}{y_p(\varepsilon)}, ..., -\frac{y_{p-1}(\varepsilon)}{y_p(\varepsilon)}, \frac{1}{y_p(\varepsilon)}, -\frac{y_{p+1}(\varepsilon)}{y_p(\varepsilon)}, ..., -\frac{y_m(\varepsilon)}{y_p(\varepsilon)}]^T$. Since the division of two Laurent series in the scalar case is not a problem [20], one can easily obtain the Laurent series for $E(\varepsilon)$

$$E(\varepsilon) = \frac{1}{\varepsilon^t} E^{(-t)} + ... + \frac{1}{\varepsilon} E^{(-1)} + E^{(0)} + ... \tag{12}$$

Consequently, the coefficients $B'^{(k)}, k = -s', -s'+1, ...$ of the Laurent series for $A_{B'}^{(-1)}(\varepsilon)$ can be calculated by the following formula

$$B'^{(k)} = \sum_{i+j=k} E^{(i)} B^{(j)}, \quad k = -s', -s'+1, ... \, . \tag{13}$$

However, we would like to emphasize that we need to update only the coefficients $B'^{(-1)}, B'^{(0)}$ by the above formula. The other coefficients, if needed, can be restored by iterative formulae (6),(7) in a more efficient way. The computational complexity for this updating procedure is $O(\bar{s}m^2)$, where $\bar{s}$ is the maximal order of poles of the Laurent expansions for basis matrices.

3 Application to Markov Decision Processes

Below we apply the asymptotic simplex method to several problems arising in MDPs with finite state and action spaces. In the following section we then extend several results to Markov games. Let $\mathbf{A}$ be the finite set of actions and let $\mathbf{I}$ be the finite set of states. In the models considered below $p(j|i,a)$ denotes the transition probability of the underlying Markov chain from state i to state j, if action a is chosen, and $r(i,a)$ denotes the corresponding immediate reward. We use the standard definitions of policies (see e.g. [9]). Let β be the initial distribution of the state (at time 1).

We consider two reward criteria below. The expected discounted reward, in which the we maximize

$$R^\alpha(u) = (1-\alpha) E_\beta^u \sum_{t=1}^{\infty} \alpha^{t-1} r(X_t, A_t),$$

where α is the discount factor, and X_t and A_t are the stochastic processes of the states and actions. In some cases below we write the discounted cost in terms of the interest factor ρ instead of the discount factor. The relation between the two is given by

$$\rho = \frac{1-\alpha}{\alpha}, \quad \alpha = \frac{1}{1+\rho}.$$

The second criterion is the expected average cost:

$$\overline{R}_\beta(u) := \liminf_{s\to\infty} \frac{1}{s} \sum_{t=1}^{s} E_\beta^u r(X_t, A_t).$$

3.1 Model I: MDP with Blackwell optimality criterion.

The Blackwell optimal policy in MDP is a discount optimal policy for all discount factor sufficiently close to one, or, equivalently, the policy which is optimal for all interest rate sufficiently close to zero [25, 26]. The Blackwell optimal policy can be determined by using the asymptotic linear programming approach [1, 8, 17]. Namely, one needs to find a solution to the following linear program which is optimal for all interest rates $\rho > 0$ and sufficiently small.

$$\max \sum_{i,a} r(i,a) x_{ia}$$

$$\sum_{i,a} [(1+\rho)\delta_{ij} - p(j|i,a)] x_{ia} = \rho \beta_j, \quad x_{ia} \geq 0, \qquad j \in \mathbf{I}. \tag{14}$$

Note that the above linear program can be immediately written in the form (1),(2) with $\varepsilon = \rho$. Hordijk *et al* [8] developed a refined version of Jeroslow's asymptotic linear programming method [14, 15] to solve (14). Later Lamond [17] applied his version of the asymptotic simplex method to the above parametric LP. Of course, our algorithm can be applied to this model as well. Lamond [17] pointed out to the fact that in this particular case the Laurent series (3) has always a simple pole. The latter implies that the complexity of the basis updating is $O(m^2)$, which is comparable with the complexity of the basis updating for the ordinary simplex method.

3.2 Model II: Markov branching decision chains.

The Markov branching decision chains were introduced by Huang and Veinott in [13]. These are MDPs with immediate rewards dependent on the interest rate. Namely, it is assumed that $r(i,a) = r^\rho(i,a)$ is a polynomial of the interest rate ρ. To find a policy which is optimal for all sufficiently small ρ

(Strong-value policy [13]), we need to solve only slightly modified version of (14), that is

$$\max \sum_{i,a} r^{\rho}(i,a)x_{ia}$$

$$\sum_{i,a}[(1+\rho)\delta_{ij} - p(j|i,a)]x_{ia} = \rho\beta_j, \quad x_{ia} \geq 0, \qquad j \in \mathbf{I}. \tag{15}$$

Note that after a simple modification of Step 2 of our algorithm [7], the asymptotic simplex method can be applied to solve this problem. The method of finding the optimal policy proposed by Huang and Veinott [13] also employs the asymptotic linear programming approach. Their method is based on the subsequent solution of the augmented LPs, whose objective functions and the right hand sides but not the constraint coefficient matrix depend on the parameter. Of course, this facilitates the implementation of the asymptotic linear programming, however they have to deal with large matrices. Moreover, since, as in the Model I, the Laurent series (3) has only a simple pole, the dependence of the constraint coefficient matrix on the parameter does not make much harm.

3.3 Model III: Singularly perturbed MDP with average criterion.

In [2] we applied the asymptotic linear programming [14, 15, 8], which is based on the ordering in the field of rational functions, to singularly perturbed MDPs with average criterion. Since one needs to solve the following parametric program

$$\max \sum_{i,a} r(i,a)x_{ia}$$

$$\sum_{i,a}[\delta_{ij} - p(j|i,a) - \varepsilon d(j|i,a)]x_{ia} = 0,$$

$$\sum_{a} x_{ja} + \sum_{i,a}[(\delta_{ij} - p(j|i,a) - \varepsilon d(j|i,a)]y_{ia} = \beta_j,$$

$$x_{ia} \geq 0, \qquad y_{ia} \geq 0, \tag{16}$$

the asymptotic simplex method, based on the Laurent series expansions, can be applied as well. Moreover, since the basis updating in the asymptotic linear programming which uses the rational function representation takes $O(m^4 log(m))$ flops, and the asymptotic simplex method which uses the Laurent series representation takes $O(\bar{s}m^2)$ flops; obviously, the latter method appears to be much faster.

3.4 Model IV: Singularly perturbed MDP with killing interest rate

Quadrat and Delebecque [5, 22] considered singularly perturbed MDPs with "killing interest rate" $\rho(\varepsilon) = \mu\varepsilon^l$, where l is the order of a time scale. This model exhibits the necessity of different control regimes for the different time scales. Quadrat and Delebecque [5, 22] provided a lexicographical policy improvement algorithm for the solution of the perturbed dynamic programming equation. Alternatively, the extension of our asymptotic simplex method for the polynomial perturbation [7] can be used in this problem. Here the parametric linear program takes the following form

$$\max \sum_{i,a} (1 + \mu\varepsilon^l) r(i,a) x_{ia}$$

$$\sum_{i,a} [(1 + \mu\varepsilon^l)\delta_{ij} - p(j|i,a) - \varepsilon d(j|i,a)] x_{ia} = 1, \quad x_{ia} \geq 0. \tag{17}$$

3.5 Generalized Model V:

Finally, we would like to note that the Models I,II and IV can be viewed as particular cases of the next unified scheme. Let transition probabilities $p^\varepsilon(j|i,a)$, immediate rewards $r^\varepsilon(i,a)$ and interest rate $\rho(\varepsilon)$ of an MDP model be polynomials of the parameter ε. Then a policy which is optimal for all sufficiently small values of parameter ε can be found from the next parametric linear program.

$$\max \sum_{i,a} (1 + \rho(\varepsilon)) r^\varepsilon(i,a) x_{ia}$$

$$\sum_{i,a} [(1 + \rho(\varepsilon))\delta_{ij} - p^\varepsilon(j|i,a)] x_{ia} = 1, \quad x_{ia} \geq 0. \tag{18}$$

The above perturbed LP can be efficiently solved by the generalization of the asymptotic simplex method for the case of polynomial dependence of the parameter [7]. Note that we retrieve Model I with $\rho(\varepsilon) = \varepsilon$, $r^\varepsilon(i,a) = r(i,a)$, $p^\varepsilon(j|i,a) = p(j|i,a)$; Model II with $\rho(\varepsilon) = \varepsilon$, $r^\varepsilon(i,a) = \sum_{k=0}^{p} \varepsilon^k r_k(i,a)$,

$p^\varepsilon(j|i,a) = p(j|i,a)$; and Model IV with $\rho(\varepsilon) = \mu\varepsilon^l$, $r^\varepsilon(i,a) = r(i,a)$, $p^\varepsilon(j|i,a) = p(j|i,a) + \varepsilon d(j|i,a)$.

3.6 Model VI: Blackwell optimality for constrained MDPs

In CMDP for each pair of state-action the immediate costs $c^k(i,a)$, $k = 1,...,K$ occur in addition to the immediate reward $r(i,a)$. The latter results in the next additional constraints:

$$W_\beta^{k,\alpha}(u) = (1-\alpha)E_\beta^u[\sum_{t=1}^{\infty} \alpha^{t-1} c^k(X_t, A_t)] \leq d_k, \tag{19}$$

where $k = 1,...,K$. Note that a policy which maximizies the expected discount reward subject to constraints (19) does not necessarily exist. Furthermore, recall that in the case of the unconstrained MDP the Blackwell optimal policy maximizes the expected discounted reward for all interest rates sufficiently small and the policy does not depend on the interest rate. By contrast, the optimal policy which maximizes the expected discount reward in the constrained MDP (if it exists) may depend on the interest rate even in an arbitrary small vicinity of zero [1]. This policy can be found from the following parametric linear program [1].

$$\max \sum_{i,a} r(i,a) x_{ia}$$

$$\sum_{i,a} [(1+\rho)\delta_{ij} - p(j|i,a)] x_{ia} = \beta_j \rho, j \in \mathbf{I},$$

$$\sum_{i,a} c^k(i,a) x_{ia} \leq d_k, \ k = 1,...,K; \quad x_{ia} \geq 0. \tag{20}$$

The above program can be solved by the asymptotic simplex method and the optimal policy $\pi^*_{ia}(\rho) = P(Action = a|State = i)$ is determined by [1, 9]

$$\pi^*_{ia}(\rho) = \begin{cases} x^*_{ia}(\rho)/\sum_a x^*_{ia}(\rho) & \text{if} \quad \sum_a x^*_{ia}(\rho) > 0, \\ \text{arbitrary} & \text{if} \quad \sum_a x^*_{ia}(\rho) = 0, \end{cases}$$

where x^* is an optimal basis solution of (20). Note that policy π^* can be randomized.

The asymptotic simplex method can thus be used for computing the optimal values and policies for constrained MDPs as well. Can this be used to provide a Blackwell optimal policy? It turns out that not only such a policy need not exist, but moreover, there need not even be an ϵ-Blackwell policy. In particular, in constrained MDPs, the limit optimal policy may be unfeaasible for all discount factors $\alpha \neq 1$. This is illustrated by the following example.

Example 1 Consider two states: 1 and 2. State 2 is absorbing: there is one action available there, and once in state 2 we always remain in that state. In state 1 there are two actions: 1 and 2, where action 1 results in remaining at state 1, and 2 results in a deterministic transition to state 2. We assume that the initial state is 1. Assume that we wish to maximize the expected discounted time spent in state 1:

$$R_1^\alpha(u) = (1-\alpha)E_1^u \sum_{t=1}^{\infty} \alpha^{t-1} r(X_t, A_t)$$

where $r(x,a) = 1\{x = 1\}$. Assume that we have a single constraint: $W_1^\alpha(u) = 1/2$, where the immediate cost related to the constraint is $c(x,a) = 1\{x = 2\}$. (Although in this example we have an equality constaint, this is a special case of our setting that can be obtained by transforming the constraint into two inequality constraints.) We know that we may restrict to stationary policies for any $\alpha < 1$ [1], where a stationary policy can be identified as the probability p of remaining in state 1 in the next move. A simple calculation gives

$$R_1^\alpha(p) = \frac{1-\alpha}{1-p\alpha}, \quad W_1^\alpha(p) = \alpha \frac{1-p}{1-\alpha p}$$

and clearly $R_1^\alpha(u) + W_1^\alpha(u) = 1$. The only feasible p obtained from the constraint $W_1^\alpha(p) = 1/2$ is

$$p = p(\alpha) = 2 - 1/\alpha.$$

Thus the limit policy $p = 1$ is not feasible for any discount factor $\alpha < 1$. Moreover, any stationary policy which is feasible for one discount factor cannot be feasible for other discount factors.

Introduce now the Slater condition: There exists some u such that for all $k = 1, ..., K$

$$J_\beta^k(u) := \limsup_{s \to \infty} \frac{1}{s} \sum_{t=1}^{s} E_\beta^u c^k(X_t, A_t) < d_k. \tag{21}$$

Then one can show [24, Thm 3.3] that there exists an ϵ-Blackwell optimal Markov policy $u(\epsilon)$, i.e. a policy which is ϵ-optimal for all α sufficiently close to 1. The ϵ-optimal policy is in general not the policy which is optimal for the limit problem, nor one that is obtained as limit (as α tends to 1) of the optimal policies for the α-discounted problems. Indeed, these two candidates need not even be feasible for the constrained MDP with any $\alpha < 1$.

To illustrate the last point consider the following

Example 2 *Consider the previous example, but replace the equality constraint by an inequality constraint: $W_1^\alpha(u) \geq 1/2$. The optimal policy is still the one given in the previous example, and the Slater condition (21) holds. The limit policy, $p = 1$, is not feasible for any $\alpha < 1$.*

4 Application to Stochastic games

An important class of zero-sum stochastic games that can be solved using a linear program has been introduced in [10, 11]. It is characterized by the fact that only one player can control the transition probabilities, but both players do control the costs. More precisely, let $\mathbf{A}$ and $\mathbf{B}$ be the finite sets of actions available to players 1 and 2, respectively, and let $\mathbf{I}$ be the finite set of states. Then $p(j|i,a)$ denotes the transition probability of the underlying Markov chain from state i to state j, if actions a and b are chosen by the players, and $r(i,a,b)$ denotes the corresponding immediate reward. We consider the standard definitions of history dependent policies for the players as in [10, 11], i.e. a policy for player i is a sequence of conditional probability distributions over the space of its actions, where the probability at time t is a function of the history of previous states and actions of both players as well as of the current state. The cost criteria are direct extensions of those we used for MDPs:

$$R^\alpha(u,v) = (1-\alpha)E_\beta^{u,v}\sum_{t=1}^{\infty}\alpha^{t-1}r(X_t,A_t,B_t),$$

$$\overline{R}_\beta(u,v) := \liminf_{s\to\infty}\frac{1}{s}\sum_{t=1}^{s}E_\beta^{u,v}r(X_t,A_t,B_t).$$

We assume that player 1 wishes to maximizes the above rewards and player 2 wishes to minimize it.

We shall also consider the case of stochastic games in which the policies of player one are conststrained by (19). In particular, the costs involved in these constraints are assumed to depend only on the actions of player 1. Constaints are analogously defined for the expected average cost.

4.1 Model VII: Constrained stochastic games with Blackwell optimality criterion.

As in the MDP case, we cannot expect there to be a Blakwell policy for the constrained stochastic game. One can show, however, that even in the absence of of the extra constraints (19) (i.e. $K = 0$) there need not be in general a Blackwell optimal policy.

To compute the values and solutions of the stochastic game we consider the following pair of dual linear programs.

$$\begin{cases} \max \sum_i z_i \\ \sum_{i,a}[(1+\rho)\delta_{ij} - p(j|i,a)]x_{ia} = \beta_j \rho, \quad j \in \mathbf{I}, \\ -\sum_a r(j,a,b)x_{ja} + z_j \leq 0, \quad b \in \mathbf{B}, j \in \mathbf{I}, \\ \sum_{i,a} c^k(i,a)x_{ia} \leq d_k, \; k = 1, ..., K, \quad x_{ja} \geq 0, \quad j \in \mathbf{I}, a \in \mathbf{A}. \end{cases} \tag{22}$$

$$\begin{cases} \min \rho \sum_i \beta_i v_i + \sum_k d_k w_k \\ (1+\rho)\sum_j(\delta_{ij} - p(j|i,a))v_j - \sum_b r(i,a,b)\eta_{ib} + \sum_k c^k(i,a)w_k \geq 0, \\ \sum_b \eta_{ib} = 1, \quad i \in \mathbf{I} \\ \eta_{ib} \geq 0, i \in \mathbf{I}, \; b \in \mathbf{B}, \quad w_k \geq 0, \; k = 1, ..., K. \end{cases} \tag{23}$$

Let (x^*, z^*) and (v^*, η^*, w^*) be the corresponding optimal solutions of the primal and dual programs, respectively. Define the decision rule π^* for player 1 as $\pi^*_{ia} = x^*_{ia} / \sum_{a'} x_{ia'}$, $a \in \mathbf{A}, i \in \mathbf{I}$ whenever the denominator is nonzero (and arbitrary otherwize).

Then (π^*, ρ^*) are saddle point stationary policies for the stochastic game. This follows directly from [10, Thm. 11].

Assume that the Slater (21) condition holds (note that this condition does not depend on the policy of player 2). Then one can show again using [24] that there exists an ϵ-Blackwell saddle point Markov pair of policies $(u(\epsilon), v(\epsilon))$. Such a policy pair is thus an ϵ-saddle point for all ρ sufficiently close to 0.

Example 3 *Consider an absorbing game with two states: I and II and action spaces* $\mathbf{A} = \{a_1, a_2\}$, $\mathbf{B} = \{b_1, b_2\}$. *State II is absorbing and the immediate reward there is constant and equals 1 (independently of the actions). When at state I, we remain there if player 1 uses* a_1, *and we move to state II if she uses* a_2; *the transition is independent of the actions of player 2.*

The immediate rewards at state I are

$$r(a_1, b_2) = r(a_2, b_1) = 1, \quad r(a_2, b_2) = r(a_1, b_1) = 0.$$

The discounted value of this game is given by $V_I = 1/(2-\alpha)$ if we start at state I and 1 if we start at state II. It can be obtained through the LP approach or directly by solving the dynamic programming:

$$V_I = val \begin{pmatrix} \alpha V_I & 1 \\ 1 - \alpha + \alpha V_I & \alpha \end{pmatrix}$$

The saddle point stationary policy for player 1 is to choose a_1 at state I with probability $p = (1-\alpha)/(2-\alpha)$ and otherwize choose a_2. The saddle point stationary policy for player 2 is to choose b_1 at state I with probability half. We thus see that the saddle point policy for player I depends on the discount factor, so that there is no Blackwell optimal policy as it was defined for unconstrained MDPs. However, the policy $p = 0$ is an ϵ-Blackwell policy for player 1 for all discount factors sufficiently close to 1.

4.2 Model VIII: Singular perturbation of stochastic games

To compute the values and solutions of the stochastic game for the case of expected average cost, we consider the following pair of dual linear programs.

$$\left\{ \begin{array}{c} \max \sum_i z_i \\ \sum_{i,a} [\delta_{ij} - p(j|i,a)] x_{ia} = 0, \quad j \in \mathbf{I}, \\ \sum_a x_{ja} + \sum_{i,a} [\delta_{ij} - p(j|i,a)] y_{ia} = \beta_j \rho, \quad j \in \mathbf{I}, \\ -\sum_a r(j,a,b) x_{ja} + z_j \leq 0, \quad b \in \mathbf{B}, j \in \mathbf{I}, \\ \sum_{i,a} c^k(i,a) x_{ia} \leq d_k, \ k = 1, ..., K, \quad x_{ja} \geq 0, y_{ja} \geq 0, \quad j \in \mathbf{I}, a \in \mathbf{A}. \end{array} \right. \tag{24}$$

$$\left\{ \begin{array}{c} \min \sum_i \beta_i v_i + \sum_k d_k w_k \\ \sum_j (\delta_{ij} - p(j|i,a)) \phi_j \geq 0 \\ \phi_j + \sum_j (\delta_{ij} - p(j|i,a)) v_j - \sum_b r(i,a,b) \eta_{ib} + \sum_k c^k(i,a) w_k \geq 0, \\ \sum_b \eta_{ib} = 1, \quad i \in \mathbf{I} \\ \eta_{ib} \geq 0, i \in \mathbf{I}, \ b \in \mathbf{B}, \quad w_k \geq 0, \ k = 1, ..., K. \end{array} \right. \tag{25}$$

We know from [11] that the values of each of the LPs are equal to that of the constrained stochastic game. Moreover, the solutions (x^*, z^*) and $(\phi^*, v^*, \eta^*, w^*)$ corresponding to the the primal and dual programs, respectively, define saddle Markov (history dependent) policies for the stochastic games. Note that even in the case of a single player, there do not exist in general optimal stationary policies for the constrained MDP.

We can now introduce the same type of perturbation as we had in Model III, where we replace the transition probabilities p by perturbed transitions $p + \epsilon d$, and then investigate the solutions at the neighborhood of $\epsilon = 0$.

References

[1] E. Altman, A. Hordijk and L.C.M. Kallenberg, "On the value function in constrained control of Markov chains", *ZOR: MMOR*, v.44, pp.387-399, 1996.

[2] E. Altman, K.E. Avrachenkov, and J.A. Filar, "Asymptotic linear programming and policy improvement for singularly perturbed Markov decision processes", *ZOR: MMOR*, v.49, pp.97-109, 1999.

[3] K. E. Avrachenkov, M. Haviv and P. G. Howlett, "The inversion of singularly perturbed matrices", submitted to *SIAM J. Matr. Anal. Appl.*, 1998.

[4] G. B. Dantzig, *Linear programming and extensions*, Princeton University Press, 1963.

[5] F. Delebecque and J.P. Quadrat, "Optimal control of Markov chains admitting strong and weak interactions", *Automatica*, v.17, pp.281-296, 1981.

[6] B. C. Eaves and U. G. Rothblum, "A theory on extending algorithms for parametric problems", *Math. Oper. Res.*, v.14, pp. 502-533, 1989.

[7] J. A. Filar, E. Altman and K. E. Avrachenkov, "An asymptotic simplex method for singularly perturbed linear programs", submitted to *Oper. Res. Letters*, 1998.

[8] A. Hordijk, R. Dekker, and L. C. M. Kallenberg, "Sensitivity-Analysis in Discounted Markovian Decision Problems", *OR Spektrum*, v.7, pp.143-151, 1985.

[9] A. Hordijk and L. C. M. Kallenberg, "Transient policies in discrete dynamic programming: Linear programming including suboptimality tests and additional constraints", *Math. Program.*, v.30, pp.46-70, 1984.

[10] A. Hordijk and L. C. M. Kallenberg, "Linear programming and Markov games I", in: O. Moeschlin and D. Pallschke (eds.), Game theory and mathematical economics, North Holland (1981) 291-305.

[11] A. Hordijk and L. C. M. Kallenberg, "Linear programming and Markov games II", in: O. Moeschlin and D. Pallaschke (eds.), Game theory and mathematical economics, North Holland (1981) 307-320.

[12] Y. Huang, “A canonical form for pencils of matrices with applications to asymptotic linear programs”, *Lin. Alg. Appl.*, v.234, pp.97-123, 1996.

[13] Y. Huang and A.F. Veinott, Jr., “Markov branching decision chains with interest-rate-dependent rewards”, *Prob. Engin. Inform. Sciences*, v.9, pp.99-121, 1995.

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

[15] R. G. Jeroslow, “Linear Programs Dependent on a Single Parameter”, *Disc. Math.*, v.6, pp.119-140, 1973.

[16] B. F. Lamond, “A generalized inverse method for asymptotic linear programming”, *Math. Program.*, v.43, pp.71-86, 1989.

[17] B. F. Lamond, "An efficient basis update for asymptotic linear programming", *Linear Algebra and Appl.*, v.184, pp.83-102, 1993.

[18] C. E. Langenhop, “The Laurent expansion for a Nearly Singular Matrix”, *Lin. Alg. Appl.*, v.4, pp.329-340, 1970.

[19] D. G. Luenberger, *Linear and nonlinear programming*, 2nd ed., Addison-Wesley, 1984.

[20] A.I. Markushevich, *Theory of functions of a complex variable*, Chelsea Publishing Company, New York, 1977.

[21] A. A. Pervozvanskii and V. G. Gaitsgori, *Theory of Suboptimal Decisions,* Kluwer Academic Publishers, Dordrecht, Netherlands, 1988.

[22] J.P. Quadrat, “Optimal control of perturbed Markov chains: the multitime scale case”, in *Singular perturbations in systems and control*, ed. M.D. Ardema, CISM Courses and Lectures no.280, Springer-Verlag, Wien - New York, 1983.

[23] P. J. Schweitzer and G. W. Stewart, “The Laurent expansion of pensils that are singular at the origin”, *Lin. Alg. Appl.*, v.183, pp.237-254, 1993.

[24] M. Tidball, O. Pourtallier and A. Altman, "Robustness of convex optimization with applications to controlled Markov chains", INRIA report No. 2933, to appear in *SIAM J. Control and Optimization*, 2000.

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

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