OR Spektrum (1985) 7:143–151

OR Spektrum

Sensitivity-Analysis in Discounted Markovian Decision Problems

A. Hordijk, R. Dekker*, and L. C. M. Kallenberg

Department of Mathematics and Computer Science, University of Leiden, P.O. Box 9512, NL-2300 RA Leiden, The Netherlands

Received 1 October 1984 / Accepted in revised form 18 July 1985

Summary. This paper deals with a finite-state, finite-action discrete-time Markov decision model. A linear programming procedure is developed for the computation of optimal policies over the entire range of the discount factor. Furthermore, a procedure is presented for the computation of a Blackwell optimal policy.

Zusammenfassung. Diese Arbeit befaßt sich mit diskreten Markoffschen Entscheidungsmodellen mit endlichen Zustands- und Aktionsräumen. Ein lineares Programm wird entwickelt für die Berechnung von optimalen Politiken über den ganzen Bereich des Diskontierungsfaktors. Anschließend wird ein Verfahren angegeben für die Bestimmung einer Blackwell-optimalen Politik.

1. Introduction

At discrete time points $t = 1, 2, \ldots$ a system is observed by a decision maker in one of the states of a finite *state-space* $E = \{1, 2, \ldots, N\}$. If, at time point t, the system is observed in state i, the decision maker controls the system by choosing an action from a finite *action set* $A(i)$, which is independent of t. If the decision maker chooses action a in state i, then the following happens independently of the history of the process:

(i) a *reward* r_{ia} is earned immediately,

(ii) the state of the system at the next time point is state j with *transition probability* p_{iaj} ($p_{iaj} \geqslant 0$, $j \in E$ and $\Sigma_j p_{iaj} = 1$).

A *decision rule* π^t at time t is a function which assigns to each action the probability of taking that action at time t; in general, it can depend on all realized states up to and including time t and on all realized actions up to time t. A *policy* R is a sequence of decision rules: $R = (\pi^1, \pi^2, \ldots, \pi^t, \ldots)$. A policy is said to be *deterministic* and *stationary* if all decision rules are identical and nonrandomized. Hence, a deterministic and stationary policy is completely described by a mapping $f: E \to \cup_i A(i)$ such that $f(i) \in A(i)$ for every $i \in E$. Such a policy is properly denoted by $(f, f, \ldots)$, however we will also write f for such a policy.

Let $\{X_t, t = 1, 2, \ldots\}$ and $\{Y_t, t = 1, 2, \ldots\}$ be the sequences of random variables, denoting the observed states and chosen actions respectively. $\mathbb{P}_R(X_t = j, Y_t = a | X_1 = i)$ denotes the probability that at time t the observed state is state j and the chosen action is action $a \in A(j)$, on condition that the state at time $t = 1$ is state i and that policy R is used. Given discount factor $\alpha \in [0, 1)$, initial state i and policy R, the *total expected discounted reward* is denoted by $v_i^\alpha(R)$, i.e.

$$v_i^\alpha(R) = \sum_{t=1}^{\infty} \alpha^{t-1} \Sigma_j \Sigma_a \mathbb{P}_R(X_t = j, Y_t = a | X_1 = i) r_{ja}. \tag{1.1}$$

Let $v_i^\alpha = \sup_R v_i^\alpha(R)$, $i \in E$. v^α is called the *value vector*. A policy R^* is called *α-optimal* if $v_i^\alpha(R^*) = v_i^\alpha$ for every $i \in E$; R^* is called *Blackwell optimal* if for some $\alpha_0 \in [0, 1)$, $v_i^\alpha(R^*) = v_i^\alpha$ for every $i \in E$ and every $\alpha \in [\alpha_0, 1)$.

We shall use linear programming and derive a simplex procedure in which the elements are not from the usual Archimedean ordered field of the real numbers, but

* Present address: Department of Mathematics and Systems Engineering, Kon./Shell Laboratory Amsterdam, P.O. Box 3003, NL-1003 AA Amsterdam, The Netherlands. The research of this author was supported by the Netherlands Foundation for Mathematics (SMC) with financial aid from the Netherlands Organization for the Advancement of Pure Research (ZWO)

from the non-Archimedean ordered field of the rational functions. The opportunity of using the simplex method for linear programs in non-Archimedean ordered fields is first observed by Charnes and Cooper [2], p. 756. Jeroslow [12] described how this concept can be used to obtain the solution of linear programs in which all coefficients are rational functions of a single parameter. In the next section, we give a description of the field $F(\mathbb{R})$ of the rational functions with real coefficients. Subsequently, we discuss the computation of optimal policies over the entire range of the discount factor. In Sect. 4, the computation of a Blackwell optimal policy is investigated.

2. The Field $F(\mathbb{R})$ of Rational Functions with Real Coefficients

Let $\mathbb{R}$ be the ordered field of the real numbers with the usual ordering denoted by $>$. By $P(\mathbb{R})$ we denote the set of all polynomials with real coefficients, i.e. the set of elements

$$p(x) = a_0 + a_1 x + \ldots + a_n x^n, \quad \text{where } a_i \in \mathbb{R},\ 1 \leqslant i \leqslant n,\ n \in \mathbb{N}. \tag{2.1}$$

By p_0 and p_1 we denote the polynomial $p_0(x) \equiv 0$ and $p_1(x) \equiv 1$ respectively. The *dominating coefficient* of a polynomial p given by (2.1) is the coefficient a_k, where k is the smallest integer with $a_k \neq 0$. The dominating coefficient of p is denoted by $\mathcal{d}(p)$. The field $F(\mathbb{R})$ of rational functions with real coefficients consists of the elements

$$\frac{p(x)}{q(x)}, \quad \text{where } p \text{ and } q \text{ are from } P(\mathbb{R}), \text{ and } q \neq p_0. \tag{2.2}$$

The polynomial p is identified with the rational function $\frac{p}{p_1}$; two rational functions $\frac{p}{q}$ and $\frac{r}{s}$ are identified (denoted by $\frac{p}{q} \,\iota = \frac{r}{s}$) if $p(x)s(x) = q(x)r(x)$ for every $x \in \mathbb{R}$. The operations + and $\cdot$ in $F(\mathbb{R})$ are the natural addition and multiplication, i.e.

$$\frac{p(x)}{q(x)} + \frac{r(x)}{s(x)} \,\iota = \frac{p(x)s(x) + r(x)q(x)}{q(x)s(x)},$$

$$\frac{p(x)}{q(x)} \cdot \frac{r(x)}{s(x)} \,\iota = \frac{p(x)r(x)}{q(x)s(x)}$$

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

$$\frac{p}{q} \,\iota > p_0 \quad \text{if and only if } \mathcal{d}(p)\mathcal{d}(q) > 0. \tag{2.3}$$

If $\frac{p}{q} \,\iota > p_0$ the rational function $\frac{p}{q}$ is called positive. $\frac{p}{q} \,\iota \geqslant p_0$ means that either $p \,\iota = p_0$ or $\frac{p}{q} \,\iota > p_0$. The fact that the above described $F(\mathbb{R})$ is a non-Archimedean ordered field can be verified quite straightforwardly (cf. Van der Waerden [17], pp. 209–210). The continuity of polynomials implies that the rational function $\frac{p}{q}$ is positive if and only if $\frac{p(x)}{q(x)} > 0$ for all x sufficiently near 0. Hence, we obtain the following result.

Lemma 2.1. *The rational function $\frac{p}{q}$ is positive if and only if there exists an $x_0 > 0$ such that $\frac{p(x)}{q(x)} > 0$ for every $x \in (0, x_0]$.*

3. Computation of Optimal Policies Over the Entire Range of Discount Factors

In this paragraph we discuss how, by linear programming, the optimal policies over the entire range of the discount factors can be computed. To this end, we first review some results from discounted dynamic programming (Sect. 3.1) and introduce the simplex format for the computations in the non-Archimedean ordered field $F(\mathbb{R})$ (Sect. 3.2). The execution of the simplex method in $F(\mathbb{R})$ requires a procedure which computes zeroes of polynomials; this is dealt with in Sect. 3.3. Then, in Sect. 3.4, we describe the simplex procedure for the computation of optimal policies over the regions of the discount factors. This procedure is illustrated by Howard's taxicab problem (Howard [10], p. 44) in Sect. 3.5. Smallwood [15] also proposed a procedure to compute the optimum policy regions. In Sect. 3.6 we compare our procedure with his approach.

3.1. Review of Discounted Dynamic Programming

Consider the Markov decision problem described in the introduction. For a deterministic and stationary policy f, $r(f)$ denotes the N-vector with i-th component $r_{if(i)}$,

and $P(f)$ is the $N \times N$ matrix with (i,j)-th element $p_{if(i)j}$. The following results are well known (e.g. Derman [3]).

Theorem 3.1.1. *(i) For a deterministic and stationary policy f, the total expected discounted rewards are given by the unique solution of the linear system*

$$x_i = r_i(f) + \alpha \Sigma_j P_{ij}(f)x_j, \quad i \in E. \tag{3.1.1}$$

(ii) The value vector v^α is the unique solution of the optimality equations

$$x_i = \max_{a \in A(i)} \{r_{ia} + \alpha \Sigma_j p_{iaj}x_j\}, \quad i \in E. \tag{3.1.2}$$

(iii) If x satisfies the system of inequalities

$$x_i \geqslant r_{ia} + \alpha \Sigma_j p_{iaj}x_j, \quad a \in A(i), i \in E, \tag{3.1.3}$$

then $x_i \geqslant v_i^\alpha$ for every $i \in E$.

(iv) For every $\alpha \in [0, 1)$ there exists a deterministic and stationary α-optimal policy f_α.

(v) There exists a deterministic and stationary Blackwell optimal policy f_*.

The discount factor α is interchangeable with the interest rate ρ. The relation between α and ρ is given by $\alpha(1+\rho) = 1$. Hence $\alpha \uparrow 1$ corresponds to $\rho \downarrow 0$ and $\alpha \downarrow 0$ corresponds to $\rho \uparrow \infty$. We shall write $v^\alpha(f)$ or $v^\rho(f)$ depending on whether the total rewards are considered as a function of α or of ρ respectively.

Equation (3.1.1) is equivalent to

$$\Sigma_j [(1+\rho)\delta_{ij} - P_{ij}(f)]x_j = (1+\rho)r_i(f), \quad i \in E, \tag{3.1.4}$$

where δ_{ij} is Kronecker's delta.

Solving (3.1.4) by Cramer's rule shows that for every $i \in E$, $v_i^\rho(f)$ is an element of $F(\mathbb{R})$, say p/q, where the degree of the polynomials p and q are N at the most.

It is well known, see e.g. Blackwell [1] and Smallwood [15], that the interval $[0, 1)$ of the discount factor can be broken down into a finite number of intervals, say $[0 = \alpha_0, \alpha_1), [\alpha_1, \alpha_2), \ldots, [\alpha_s, \alpha_{s+1} = 1)$, in such a way that there exist deterministic and stationary policies f_i, $0 \leqslant i \leqslant s$, where f_i is α-optimal for all $\alpha \in [\alpha_i, \alpha_{i+1})$. The number s varies with the data of the problem. Hence, on any interval the components of the value vector v^ρ are elements of $F(\mathbb{R})$.

Furthermore, Eq. (3.1.2) implies that v^ρ satisfies

$$(1+\rho)v_i^\rho \geqslant (1+\rho)r_{ia} + \Sigma_j p_{iaj}v_j^\rho, \quad a \in A(i), i \in E, \rho > 0. \tag{3.1.5}$$

Therefore, in the ordered field $F(\mathbb{R})$, we have

$$(1+\rho) \cdot v_i^\rho \,{}_l\!\geqslant (1+\rho)r_{ia} + \Sigma_j p_{iaj}v_j^\rho, \quad a \in A(i), i \in E. \tag{3.1.6}$$

In general v_i^ρ is not an element of $F(\mathbb{R})$, but there are elements of $F(\mathbb{R})$ coinciding piecewise with v_i^ρ. In (3.1.6) the components v_i^ρ, $i \in E$, have to be considered as the elements of $F(\mathbb{R})$ coinciding with v_i^ρ, $i \in E$, on the interval $[\alpha_s, \alpha_{s+1} = 1)$.

An N-vector $w(\rho)$ with components in $F(\mathbb{R})$ is called *superharmonic* if

$$(1+\rho) \cdot w_i(\rho) \,{}_l\!\geqslant (1+\rho)r_{ia} + \Sigma_j p_{iaj}w_j(\rho), \quad a \in A(i), i \in E. \tag{3.1.7}$$

The concept of superharmonicity is useful to derive linear programs for stochastic dynamic programming problems (cf. Hordijk [6], Kallenberg [13], Hordijk and Kallenberg [8], [9]).

Lemma 3.1.2. *If $w(\rho)$ is superharmonic, then $w_i(\rho) \,{}_l\!\geqslant v_i^\rho$ for every $i \in E$.*

Proof. Since $w(\rho)$ is superharmonic and as there are only a finite number of states and actions, there exists a $\rho_1 > 0$ such that

$$(1+\rho)w_i(\rho) \geqslant (1+\rho)r_{ia} + \Sigma_j p_{iaj}w_j(\rho), \quad a \in A(i), i \in E, \rho \in (0, \rho_1].$$

Hence, for every $\rho \in (0, \rho_1]$, $w(\rho)$ satisfies Eq. (3.1.5) which is equivalent to Eq. (3.1.3). Therefore, by Theorem 3.1.1(iii),

$$w_i(\rho) \geqslant v_i^\rho \quad \text{for every } i \in E \text{ and } \rho \in (0, \rho_1], \quad \text{i.e.}$$

$$w_i(\rho) \,{}_l\!> v_i^\rho \text{ for every } i \in E. \qquad \square$$

3.2. The Simplex Format

Lemma 3.1.2 implies that the value-vector v^ρ of the interval $(0, \rho,]$ can be found as the optimal solution to the following linear program in $F(\mathbb{R})$:

$$\min \{\Sigma_j w_j(\rho) \mid \Sigma_j [(1+\rho)\delta_{ij} - p_{iaj}] \cdot w_j(\rho) \,{}_l\!\geqslant (1+\rho)r_{ia}, a \in A(i), i \in E\}. \tag{3.2.1}$$

Consider also the following linear program in $F(\mathbb{R})$, which is called the *dual program* of (3.2.1):

$$\max \left\{ \Sigma_i \Sigma_a r_{ia}(1+\rho) \cdot x_{ia}(\rho) \;\middle|\; \begin{array}{l} \Sigma_i \Sigma_a [(1+\rho)\delta_{ij} - p_{iaj}] \cdot x_{ia}(\rho)\ {}_i = p_1, j \in E; \\ x_{ia}(\rho)\ {}_i \geqslant p_0, a \in A(i), i \in E. \end{array} \right\}. \qquad (3.2.2)$$

Remark. For a fixed real value of ρ the linear programs (3.2.1) and (3.2.2) are precisely the linear programs introduced by d'Epenoux [4] to compute a ρ-optimal policy. It is well known that there is a one-to-one correspondence between the extreme points of (3.2.2) and the set of deterministic and stationary policies (e.g. Derman [3]). Furthermore, for each deterministic and stationary policy f the corresponding point satisfies $x_{if(i)}(\rho) > 0$ for all $\rho > 0$.

In the sequel we will, as is also the case in the simplex method, rewrite the equalities

$$\Sigma_i \Sigma_a [(1+\rho)\delta_{ij} - p_{iaj}] \cdot x_{ia}(\rho) = 1, \quad j \in E,$$

such that at any iteration there is precisely one positive $x(\rho)$ component in each state. The only difference with the normal simplex method for a fixed value of ρ is that instead of real numbers, the elements are rational functions. In order to understand the following paragraphs the reader should be familiar with linear programming concepts as "basic solution", "complementary slackness", "condensed simplex tableau" and "reduced costs". These concepts can be found in the standard textbooks on linear programming.

As in the simplex method, for any interation, the set of constraints is written in a special way:

$$x_B = B^{-1}e - B^{-1}Nx_N, \qquad (3.2.3)$$

where e is the vector with all the components equal to p_1, x_B and x_N are the basic and nonbasic variables, B is the basic matrix and N consists of the remaining columns. We shall solve program (3.2.2) in such a way that the optimality of some basic solution, or equivalently some deterministic and stationary policy, is shown on a certain interval for the value of ρ. This is possible, because for every fixed ρ in that interval the corresponding simplex tableau is an optimal one (cf. D'Epenoux [4]). At any iteration of the simplex method there is a feasible solution $x(\rho)$ of (3.2.2) and a trial solution $w(\rho)$ of (3.2.1), called the reduced cost vector, such that the complementary slackness conditions hold, i.e.

$$x_{ia}(\rho) \cdot \{\Sigma_j [(1+\rho)\delta_{ij} - p_{iaj}] \cdot w_j(\rho) - (1+\rho)r_{ia}\} = 0,$$
$$a \in A(i), i \in E, \qquad (3.2.4)$$

for all ρ in the interval which is considered. Since any basic solution corresponds to a deterministic and stationary policy, in each state i there is exactly one action, say action $f(i)$, such that $x_{if(i)}(\rho) > 0$ for all ρ in the actual interval. Hence, by (3.2.4),

$$\Sigma_j [(1+\rho)\delta_{ij} - P_{ij}(f)] \cdot w_j(\rho) = (1+\rho)r_i(f), \quad i \in E,$$

for all actual ρ. From Theorem 3.1.1(i) it follows that $w(\rho) = v^\rho(f)$ in the actual interval.

The organization of the simplex format is based on the following lemma.

Lemma 3.2.1. *i) The elements of the simplex tableau can be written as rational functions with the same denominator, which is the product of the previous pivot elements.*

ii) The numerator and denominator of the rational functions are polynomials with degree N at the most, except for the reduced costs where the numerator can have degree $N+1$.

iii) For ρ sufficiently large, the optimal solution $x(\rho)$ is given by the basic variables $x_{if(i)}(\rho)$, where $f(i)$ is such that $r_{if(i)} = \max_{a \in A(i)} r_{ia}$, $i \in E$.

iv) The pivot operations in the simplex tableau are as follows ($n(\rho)$ is the common denominator)

a) The numerator of the pivot becomes the next common denominator, and the last common denominator becomes the new numerator of the pivot.

b) The numerators of the other elements in the pivot row are unchanged; the numerators of the other elements in the pivot column are multipled by -1.

c) For the other elements, say numerator $p(\rho)$, we replace $p(\rho)$ by $\dfrac{p(\rho)q(\rho) - r(\rho)s(\rho)}{n(\rho)}$, which is a polynomial, where $q(\rho)$ is the numerator of the old pivot, $r(\rho)$ is the numerator of the pivot row which is in the same column as $p(\rho)$, and $s(\rho)$ is the numerator in the pivot column which is the same row as $p(\rho)$.

Proof. i) Since the constraints of (3.2.2) are equalities, artificial variables $z_j(\rho)$, $j \in E$, are introduced one for each constraint. Starting the simplex method, the first basic matrix B is the identity matrix I corresponding to the artificial variables. Hence, in the first simplex tableau the elements are polynomials (in fact linear functions) of ρ, i.e. rational functions with common denominator 1. It is well known from the theory of linear programming (e.g. Zoutendijk [18]) that the elements in the simplex tableau can be written with the determinant of the basis matrix, i.e. the product of all previous pivot elements, as common denominator. This result, with a similar proof, is also valid for linear programming in $F(\mathbb{R})$.

ii) Any basic matrix is of the form $(1+\rho)I - P(f)$, i.e. it has linear functions of ρ on the diagonal and constants on the off-diagonal. Hence, by Cramer's rule, the elements of the inverse basis matrix are rational functions whose numerator is a polynomial of degree $N-1$ at the most and with a polynomial of degree N at the most as denominator. From Eq. (3.4.1) it follows that the elements in the simplex tableau have a polynomial with degree N at the most as numerator. Since in the objective function (see (3.2.2)) the variables are multiplied by a linear function, the reduced costs may have a numerator with degree $N+1$ at the most.

iii) Since $\rho \uparrow \infty$ corresponds to $\alpha \downarrow 0$, the rewards in the first period dominate the rewards earned later on. Hence, the given variables are optimal.

iv) Consider the following elements of the simplex tableau:

pivot row

$\frac{t(\rho)}{n(\rho)}$	$\frac{r(\rho)}{n(\rho)}$
$\frac{s(\rho)}{n(\rho)}$	$\frac{p(\rho)}{n(\rho)}$

pivot column

This tableau is transformed by the usual rules into:

$\frac{n(\rho)}{t(\rho)}$	$\frac{r(\rho)}{t(\rho)}$
$-\frac{s(\rho)}{t(\rho)}$	$\frac{p(\rho)\cdot q(\rho) - r(\rho)\cdot s(\rho)}{n(\rho)\cdot t(\rho)}$

(3.2.5)

Since the next common denominator is $n(\rho)\cdot\frac{t(\rho)}{n(\rho)} = t(\rho)$ (see part (i) of this Lemma) $\frac{p(\rho)\cdot q(\rho) - r(\rho)\cdot s(\rho)}{n(\rho)}$ is a polynomial and (3.2.5) proves this part of the lemma. □

Remark. Starting with the artificial variables z_j, $j \in E$ as basic variables, we can compute the optimal simplex tableau for $\alpha = 0$, or equivalently $\rho = \infty$, by exchanging $x_{1f(1)}$ with $z_1, \ldots, x_{Nf(N)}$ with z_N, where $f(i)$ is such that $\max_{a \in A(i)} r_{ia} = r_{if(i)}$, $1 \leqslant i \leqslant N$.

This tableau is optimal for $\rho \geqslant \rho_1$, where ρ_1 is the smallest value such that the reduced costs are nonnegative. To compute ρ_1 we have to compute the zeroes of some polynomials. This is the subject of the following section.

3.3 The Computation of Optimal Intervals

For any simplex tableau the interval $[\rho_0, \rho_1]$ has to be determined so that for $\rho \in [\rho_0, \rho_1]$ the corresponding deterministic and stationary policy is a ρ-optimal one. Therefore, the reduced costs have to be nonnegative in this interval.

Since for any interest rate ρ, every deterministic and stationary policy corresponds to a basic solution, the pivot elements are positive for every ρ. Consequently, the common denominators of the tableaux are also positive for every ρ. In addition, as the first feasible tableau is optimal for ρ sufficiently large, i.e. on the interval $[\rho_0, \infty)$, we have tableaux for which the upper bound of the interval is known. Hence, for the polynomials corresponding to the numerators of the reduced costs the smallest ρ_0 has to be determined such that these polynomials are nonnegative on $[\rho_0, \rho_1]$.

The computation of ρ_0 can be carried out by means of a numerical method, e.g. the method based on Sturm's Theorem (see below). With this theorem, the number $n(g, a, b)$ of real roots of a polynomial g in the interval $[a, b]$ can be determined.

Let $g_1, g_2, \ldots, g_K$ be the numerators of the reduced costs corresponding to the nonbasic variables $x_{ia}(\rho)$. Hence, $K = \Sigma_i \#A(i) - N$. The computation of ρ_0 can then be carried out by the following algorithm.

Step 0: $\rho_0 := 0$; $k := 0$; choose $\epsilon > 0$.

Step 1: $k := k+1$; if $k > K$, then stop.

Step 2: If $n(g_k, \rho_0, \rho_1) \geqslant 1$, then $\rho_2 := \frac{1}{2}(\rho_0 + \rho_1)$, $\rho_3 := \rho_1$ and go to Step 3. Otherwise, go to Step 1.

Step 3: If $n(g_k, \rho_2, \rho_3) = 0$, then $\rho_3 := \rho_2$, $\rho_2 := \frac{1}{2}(\rho_0 + \rho_2)$ and go to Step 4. Otherwise, $\rho_0 := \rho_2$, $\rho_2 := \frac{1}{2}(\rho_2 + \rho_3)$ and go to Step 4.

Step 4: If $|\rho_3 - \rho_2| \leqslant \epsilon$, then go to Step 1; otherwise, go to Step 3.

The computation of $n(g, a, b)$ is given by Sturm's Theorem (cf. Stoer and Bulirsch [16] p. 281 and Van der Waerden [17] p. 220).

Sturm's Theorem. *Let $g^{(1)}$ be the derivative of g and let $g^{(2)}, g^{(3)}, \ldots, g^{(r)}$ be determined by the Euclidean algorithm, i.e.*

$$g = h_1 g^{(1)} - g^{(2)}, \quad \text{where degree } g^{(2)} < \text{degree } g^{(1)}$$
$$g^{(1)} = h_2 g^{(2)} - g^{(3)}, \quad \text{where degree } g^{(3)} < \text{degree } g^{(2)}$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$g^{(r-2)} = h_{r-1} g^{(r-1)} - g^{(r)}, \quad \text{where degree } g^{(r)} < \text{degree } g^{(r-1)}$$
$$g^{(r-1)} = h_r g^{(r)}, \quad \text{and } h_i \text{ is a polynomial } 1 \leqslant i \leqslant r.$$

Let $\sigma(\rho)$ be the variations in sign in the number sequence $g(\rho), g^{(1)}(\rho), \ldots, g^{(r)}(\rho)$ in which all zeroes are omitted. If a and b are any numbers $(a < b)$ for which $g(a)g(b) \neq 0$, then $n(g, a, b) = \sigma(a) - \sigma(b)$, where multiple roots are counted only once.

Hence, for any simplex tableau, the interval $[\rho_0, \rho_1]$ can be computed such that this simplex tableau is an optimal tableau for every $\rho \in [\rho_0, \rho_1]$. Furthermore, let ρ_0 be obtained by the polynomial corresponding to the nonbasic variable x_{ka_k}. Then x_{ka_k} is the variable which becomes basic for the next tableau. This variable has to be exchanged with the basic variables x_{ka} (this is a unique variable).

3.4. Description of the Algorithm

By adding artificial variables $z_j(\rho), j \in E$, the linear system becomes

$$\Sigma_i \Sigma_a [(1+\rho)\delta_{ij} - p_{iaj}] \cdot x_{ia}(\rho) + z_j(\rho) = 1, \quad j \in E$$

$$x_{ia}(\rho) \geqslant 0, \quad a \in A(i), i \in E.$$

Start with the simplex tableau in which $z_j(\rho), j \in E$, are the basic variables. Then, for $j = 1, 2, \ldots, N$ exchange $z_j(\rho)$ with $x_{ja_j}(\rho)$, where a_j is such that $r_{ja_j} = \max_{a \in A(j)} r_{ja}$. The corresponding tableau is optimal for ρ sufficiently large. Next, by Sturm's Theorem, ρ_0 is determined in such a way that the present tableau is optimal for every $\rho \geqslant \rho_0$. The computation of ρ_0 determines the next basic variables. Hence, the next simplex tableau can be formed, and $\rho_1 := \rho_0$. Again by Sturm's Theorem, ρ_0 is determined so that the tableau is optimal for $\rho \in [\rho_0, \rho_1]$ and so on. The procedure terminates when $\rho_0 = 0$.

3.5. An Example

Consider the following example which is taken from Howard [10] p. 44.

State i	Action a	Transition probabilities p_{iaj}			Reward r_{ia}
		$j=1$	$j=2$	$j=3$	
1	1	1/2	1/4	1/4	8
	2	1/16	3/4	3/16	11/4
2	1	1/2	0	1/2	16
	2	1/16	7/8	1/16	15
3	1	1/4	1/4	1/2	7
	2	1/8	3/4	1/8	4

For this example the objective function becomes:

$$(1+\rho) \cdot \{8x_{11}(\rho) + \frac{11}{4}x_{12}(\rho) + 16x_{21}(\rho) + 15x_{22}(\rho) + 7x_{31}(\rho) + 4x_{32}(\rho)\}$$

and the linear system is

$$\left(\frac{1}{2}+\rho\right) \cdot x_{11}(\rho) + \left(\frac{15}{16}+\rho\right) \cdot x_{12}(\rho) - \frac{1}{2}x_{21}(\rho) - \frac{1}{16}x_{22}(\rho) - \frac{1}{4}x_{31}(\rho) - \frac{1}{8}x_{32}(\rho) = 1,$$

$$-\frac{1}{4}x_{11}(\rho) - \frac{3}{4}x_{12}(\rho) + (1+\rho) \cdot x_{21}(\rho) + \left(\frac{1}{8}+\rho\right) \cdot x_{22}(\rho) - \frac{1}{4}x_{31}(\rho) - \frac{3}{4}x_{32}(\rho) = 1,$$

$$-\frac{1}{4}x_{11}(\rho) - \frac{3}{16}x_{12}(\rho) - \frac{1}{2}x_{21}(\rho) - \frac{1}{16}x_{22}(\rho) + \left(\frac{1}{2}+\rho\right) \cdot x_{31}(\rho) + \left(\frac{7}{8}+\rho\right) \cdot x_{32}(\rho) = 1,$$

$$x_{11}(\rho), x_{12}(\rho), x_{21}(\rho), x_{22}(\rho), x_{31}(\rho), x_{32}(\rho) \geqslant 0.$$

The corresponding first simplex tableau is:

	1	$x_{11}(\rho)$	$x_{12}(\rho)$	$x_{21}(\rho)$	$x_{22}(\rho)$	$x_{31}(\rho)$	$x_{32}(\rho)$
$z_1(\rho)$	1	$\underline{\frac{1}{2}+\rho}$	$\frac{15}{16}+\rho$	$-\frac{1}{2}$	$-\frac{1}{16}$	$-\frac{1}{4}$	$-\frac{1}{8}$
$z_2(\rho)$	1	$-\frac{1}{4}$	$-\frac{3}{4}$	$1+\rho$	$\frac{1}{8}+\rho$	$-\frac{1}{4}$	$-\frac{3}{4}$
$z_3(\rho)$	1	$-\frac{1}{4}$	$-\frac{3}{16}$	$-\frac{1}{2}$	$-\frac{1}{16}$	$\frac{1}{2}+\rho$	$\frac{7}{8}+$
	0	$-8-8\rho$	$-\frac{11}{4}-\frac{11}{4}\rho$	$-16-16\rho$	$-15-15\rho$	$-7-7\rho$	$-4-4$

The common denominator is the top left-hand element in the tableau and the underlined element is the pivot element. Using the transformation described in Sect. 3.2 the next tableau is obtained:

	$\frac{1}{2}+\rho$	$z_1(\rho)$	$x_{12}(\rho)$	$x_{21}(\rho)$	$x_{22}(\rho)$	$x_{31}(\rho)$	$x_{32}(\rho)$
$x_{11}(\rho)$	1	1	$\frac{15}{16}+\rho$	$-\frac{1}{2}$	$-\frac{1}{16}$	$-\frac{1}{4}$	$-\frac{1}{8}$
$z_2(\rho)$	$\frac{3}{4}+\rho$	$\frac{1}{4}$	$\frac{-9}{64}-\frac{1}{2}\rho$	$\frac{3}{8}+\frac{3}{2}\rho+\rho^2$	$\frac{3}{64}+\frac{5}{8}\rho+\rho^2$	$-\frac{3}{16}-\frac{1}{4}\rho$	$-\frac{13}{32}-\frac{3}{4}\rho$
$z_3(\rho)$	$\frac{3}{4}+\rho$	$\frac{1}{4}$	$\frac{9}{64}+\frac{1}{16}\rho$	$-\frac{3}{8}-\frac{1}{2}\rho$	$-\frac{3}{64}-\frac{1}{16}\rho$	$\frac{3}{16}+\rho+\rho^2$	$\frac{13}{32}+\frac{11}{8}\rho+\rho^2$
	$8+8\rho$	$8+8\rho$	$\frac{49}{8}+\frac{91}{8}\rho+\frac{21}{4}\rho^2$	$-12-28\rho-16\rho^2$	$-8-23\rho-15\rho^2$	$-\frac{11}{2}-\frac{25}{2}\rho-7\rho^2$	$-3-7\rho-4\rho^2$

After inserting $x_{21}(\rho)$ and $x_{31}(\rho)$ into the basis, the first feasible tableau is reached:

	$\frac{15}{16}\rho+2\rho^2+\rho^3$	$z_1(\rho)$	$x_{12}(\rho)$	$z_2(\rho)$	$x_{22}(\rho)$	$z_3(\rho)$	$x_{32}(\rho)$
$x_{11}(\rho)$	$\frac{9}{8}+\frac{9}{4}\rho+\rho^2$	$\frac{3}{8}+\frac{3}{2}\rho+\rho^2$	$\frac{87}{64}\rho+\frac{39}{16}\rho^2+\rho^3$	$\frac{3}{8}+\frac{1}{2}\rho$	$\frac{21}{64}\rho+\frac{7}{16}\rho^2$	$\frac{3}{8}+\frac{1}{4}\rho$	$\frac{1}{32}\rho+\frac{1}{8}\rho^2$
$x_{21}(\rho)$	$\frac{9}{16}+\frac{3}{2}\rho+\rho^2$	$\frac{3}{16}+\frac{1}{4}\rho$	$-\frac{3}{8}\rho-\frac{1}{2}\rho^2$	$\frac{3}{16}+\rho+\rho^2$	$\frac{9}{32}\rho+\frac{9}{8}\rho^2+\rho^3$	$\frac{3}{16}+\frac{1}{4}\rho$	$-\frac{3}{8}\rho-\frac{1}{2}\rho^2$
$x_{31}(\rho)$	$\frac{9}{8}+\frac{9}{4}\rho+\rho^2$	$\frac{3}{8}+\frac{1}{4}\rho$	$-\frac{3}{64}\rho+\frac{1}{16}\rho^2$	$\frac{3}{8}+\frac{1}{2}\rho$	$\frac{21}{64}\rho+\frac{7}{16}\rho^2$	$\frac{3}{8}+\frac{3}{2}\rho+\rho^2$	$\frac{41}{32}\rho+\frac{19}{8}\rho^2+\rho^3$
	$\frac{207}{8}+\frac{669}{8}\rho+\frac{355}{4}\rho^2+31\rho^3$	$\frac{69}{8}+\frac{211}{8}\rho+\frac{103}{4}\rho^2+8\rho^3$	$\frac{63}{32}\rho+\frac{269}{32}\rho^2+\frac{187}{16}\rho^3+\frac{21}{4}\rho^4$	$\frac{69}{8}+\frac{257}{8}\rho+\frac{79}{2}\rho^2+16\rho^3$	$-\frac{297}{64}\rho-\frac{645}{64}\rho^2-\frac{71}{16}\rho^3+\rho^4$	$\frac{69}{8}+\frac{201}{8}\rho+\frac{47}{2}\rho^2+7\rho^3$	$-\frac{17}{32}\rho+\frac{35}{32}\rho^2+\frac{37}{8}\rho^3+3\rho^4$

Using Sturm's Theorem and the algorithm of Sect. 3.3, it follows that this tableau is optimal for $\rho \in [6.18, \infty)$. This interval for the interest rate corresponds to the interval $[0, 0.14]$ for the discount factor α.

For $\rho < 6.18$, $x_{22}(\rho)$ becomes the next basic variable. After one transformation, in which $x_{22}(\rho)$ and $x_{21}(\rho)$ are exchanged, the optimal interval for the next tableau can be computed. It turns out that this simplex tableau is optimal for $\rho \in [0.91, 6.18]$, or equivalently $\alpha \in [0.14, 0.52]$. Thereafter, the new basic variable becomes $x_{32}(\rho)$. The corresponding tableau is optimal for $\rho \in [0.27, 0.91]$, i.e. $\alpha \in [0.52, 0.79]$. Finally, the new basis variable becomes $x_{12}(\rho)$ and the corresponding tableau (see below) is optimal for $\rho \in (0, 0.27]$, i.e. $\alpha \in [0.79, 1)$. Therefore, the policy f where $f(i) = 2$, $i \in E$, is Blackwell optimal.

	$\frac{119}{128}\rho+\frac{31}{16}\rho^2+\rho^3$	$z_1(\rho)$	$x_{11}(\rho)$	$z_2(\rho)$	$x_{21}(\rho)$	$z_3(\rho)$	$x_{32}(\rho)$
$x_{12}(\rho)$	$\frac{3}{16}+\frac{19}{16}\rho+\rho^2$	$\frac{1}{16}+\rho+\rho^2$	$\frac{33}{64}\rho+\frac{3}{2}\rho^2+\rho^3$	$\frac{1}{16}+\frac{1}{16}\rho$	$-\frac{7}{16}\rho-\frac{7}{16}\rho^2$	$\frac{1}{16}+\frac{1}{8}\rho$	$-\frac{9}{64}\rho-\frac{1}{8}\rho^2$
$x_{22}(\rho)$	$\frac{153}{64}+\frac{53}{16}\rho+\rho^2$	$\frac{51}{64}+\frac{3}{4}\rho$	$\frac{17}{32}\rho+\frac{1}{2}\rho^2$	$\frac{51}{64}+\frac{29}{16}\rho+\rho^2$	$\frac{119}{64}\rho+\frac{45}{16}\rho^2+\rho^3$	$\frac{51}{64}+\frac{3}{4}\rho$	$\frac{17}{32}\rho+\frac{1}{2}\rho^2$
$x_{32}(\rho)$	$\frac{27}{128}+\frac{21}{16}\rho+\rho^2$	$\frac{9}{128}+\frac{3}{16}\rho$	$-\frac{15}{128}\rho-\frac{1}{16}\rho^2$	$\frac{9}{128}+\frac{1}{16}\rho$	$-\frac{63}{128}\rho-\frac{7}{16}\rho^2$	$\frac{9}{128}+\frac{17}{16}\rho+\rho^2$	$\frac{69}{128}\rho+\frac{25}{16}\rho^2+\rho^3$
	$\frac{1191}{32}+\frac{6107}{64}\rho+\frac{5117}{64}\rho^2+\frac{87}{4}\rho^3$	$\frac{397}{32}+\frac{869}{32}\rho+\frac{35}{2}\rho^2+\frac{11}{4}\rho^3$	$\frac{379}{256}\rho-\frac{677}{256}\rho^2-\frac{75}{8}\rho^3-\frac{21}{4}\rho^4$	$\frac{397}{32}+\frac{2561}{64}\rho+\frac{2727}{64}\rho^2+15\rho^3$	$\frac{315}{32}+\frac{1157}{64}\rho^2+\frac{463}{64}\rho^3-\rho^4$	$\frac{397}{32}+\frac{113}{4}\rho+\frac{635}{32}\rho^2+4\rho^3$	$\frac{827}{256}\rho+\frac{787}{256}\rho^2-\frac{101}{32}\rho^3-3\rho^4$

3.6. Comparison with Smallwood's Method

With respect to the complexity of our approach, we make the following observations:

A. In order to compute a new element in the simplex tableau we have to carry out the following operations (cf. Lemma 3.2.1(iv)c):

i) Two multiplications of two polynomials of degree $\mathcal{O}(N) : \mathcal{O}(N^2)$.

ii) One subtraction of two polynomials of degree $\mathcal{O}(N) : \mathcal{O}(N)$.

iii) One division of two polynomials of degree $\mathcal{O}(N) : \mathcal{O}(N^2)$.

Hence, the computation of a new element is of order N^2 and the computation of a new column is $\mathcal{O}(N^3)$. Let $A = \sum_{i=1}^{N} \#A(i)$, then the simplex tableau has A columns. Therefore, the computation of a new tableau

is $\mathcal{O}(AN^3)$. To compute the first feasible tableau (corresponding to $\alpha = 0$), N transformations have to take place, which means that this part of the procedure is $\mathcal{O}(AN^4)$.

B. For a given polynomial of degree $\mathcal{O}(N)$, the analysis of Sturm's Theorem is of order $N^2({}^2\log[(\rho_1 - \rho_0)/\epsilon])$, where ρ_0, ρ_1 and ϵ are described in Sect. 3.3, namely:

i) The construction of $g^{(j)}$ from $g^{(j-1)}$ and $g^{(j-2)}$ is of order N(degree $g^{(j-2)}$ – degree $g^{(j-1)}$. Hence the Euclidian algorithm is of order N^2.

ii) Using Horner's scheme (cf. Stoer and Bulirsch [16] p. 270) the computation of $n(g_k, \rho_2, \rho_3)$ is of order N^2.

iii) Since bisection is used, after $[{}^2\log(\rho_1 - \rho_0)/\epsilon]$ iterations the original interval $[\rho_0, \rho_1]$ is reduced to an interval of length ϵ at the most.

As in the worst case for any simplex tableau, $\mathcal{O}(A)$ analyses of the above type will be necessary, the corresponding computations are $\mathcal{O}(AN^2[{}^2\log(\rho_1 - \rho_0)/\epsilon])$. Let $\bar{\rho}$ be such that the first feasible tableau is optimal for $\rho \geqslant \bar{\rho}$, and let the problem be such that there are M optimal intervals. Then, in addition to the simplex transformations, $\mathcal{O}(AMN^2[{}^2\log\bar{\rho}/\epsilon])$ computations are required. Hence, the overall complexity of the algorithm is $\mathcal{O}(AN^2\{N^2 + M({}^2\log\bar{\rho}/\epsilon + N)\})$.

Smallwood [15] described a method to find the optimal policies in a fixed interval $[\alpha_1, \alpha_2]$, where $0 \leqslant \alpha_1 < \alpha_2 < 1$. Thus the difference between his method and ours is that we can determine a Blackwell optimal policy and he can not. With respect to the complexity of the calculations, Smallwood's method requires in any optimal interval:

i) The computation of the coefficients of a characteristic polynomial of a $N \times N$ matrix: $\mathcal{O}(N^3)$ by Danilevsky's method.

ii) The computation of some coefficients: $\mathcal{O}(N^3)$.

iii) For every policy which deviates from a fixed policy in only one action, some coefficients are calculated and an analysis by Sturm's Theorem is carried out: $\mathcal{O}(A\{N^2 + N^2[{}^2\log 1/\epsilon]\})$.

Let M be the number of optimal intervals, then the complexity of Smallwood's approach is $\mathcal{O}(MN^2\{N + A^2 \log 1/\epsilon\})$.

4. Computation of a Blackwell Optimal Policy

The method described in the previous section terminates with a Blackwell optimal policy. However, if we are only interested in the computation of a Blackwell optimal policy, we can skip the calculation of the intervals. The method based on linear programming can then be stated as follows:

1. Start with any deterministic and stationary policy and compute the corresponding simplex tableau (this is similar to the setup of the first feasible tableau in Sect. 3).

2. If every reduced cost is nonnegative with respect to the ordering in $F(\mathbb{R})$, – i.e. the dominating coefficient of the numerator of any reduced cost is nonnegative –, then the corresponding policy is Blackwell optimal. Otherwise: go to Step 3.

3. a) Take any column with a negative reduced cost as pivot column.

b) Executive one pivot transformation.

c) Go to Step 2.

Remarks. i) In our method, a sequence of tableaus is produced. Since in any transformation, the value of the objective function strictly increases (there is no degeneration), none of the bases can return. Since there are a finite number of bases, the method is finite. Furthermore, the final tableau has the property of being optimal for ρ near enough to zero. Hence, the corresponding policy is Blackwell optimal.

ii) As shown in Sect. 3.6 the number of elementary operations in one pivot step is $\mathcal{O}(AN^3)$.

iii) In the final tableau, we can compute a ρ_1 by Sturm's Theorem (as described in Sect. 3.3), such that the Blackwell optimal policy is ρ-optimal for every $\rho \in (0, \rho_1]$.

iv) Jeroslow [11] presented a policy improvement algorithm with computations in the field $F(\mathbb{R})$ of the rational functions. Our approach is the linear programming pendant; as is the case for discounted dynamic programming with a fixed discount factor, both approaches are equivalent.

v) The first method of computing a Blackwell optimal policy is derived from Miller and Veinott [14]. They proposed (in the worst case) an N-step procedure. The number of operations in every step is dominated by the solution to a system of linear equations. In the k-th step the dimension of the system is $\mathcal{O}(kN)$. Hence, the complexity of Miller and Veinott's method is $\mathcal{O}(N^7)$.

vi) The main results of this paper are obtained in 1981 and reported in [7]. Holzbaur has embroidered on our work. During the period 1981–1984 he has investigated decision problems over ordered fields. He has obtained optimality and sensitivity results for this general class of problems, reported in his interesting dissertation [6].

vii) In Step 1 of the algorithm of Sect. 4 it would seem worthwhile to start with a ρ-optimal policy for some small value of ρ.

viii) The size of the problems for which our approach is applicable in practice depends on the computer available. We expect that on a mainframe moderated sized problems can be solved. In order to get an impression, consider Howard's automobile replacement problem ([10] p. 90). The corresponding simplex tableau has 40 rows, 800 variables and 32,000 elements and each element is a polynomial of degree 40. Although this problem can be handled presumably, it is preferable to use advanced techniques based on the structure of the problem.

Acknowledgement. The authors would like to thank the referees for their helpful comments.

References

1. Blackwell D (1962) Discrete dynamic programming. Ann Math Statist 33:719–726
2. Charnes A, Cooper WW (1961) Management models and industrial applications of linear programming, vols 1 and 2. John Wiley, New York
3. Derman C (1970) Finite state Markovian decision processes. Academic Press, New York
4. D'Epenoux F (1960) Sur un probleme de production et de stockage dans l'aléatoire. Rev Fr Inform Rech Oper 14: 3–16 (Engl transl Mang Sci 10:98–108)
5. Holzbaur UD (1984) Entscheidungsmodelle über angeordneten Körpern. PhD Thesis, Universität Ulm
6. Hordijk A (1974) Dynamic programming and Markov potential theory. Math Centre Tract No 51 (Amsterdam)
7. Hordijk A, Dekker R, Kallenberg LCM (1981) A simplex-like algorithm to compute a Blackwell optimal policy. Report No 81–37 of the Inst Appl Math Comp Sci (University of Leiden)
8. Hordijk A, Kallenberg LCM (1984) Constrained undiscounted stochastic dynamic programming. Math Oper Res 9:276–289
9. Hordijk A, Kallenberg LCM (1984) Transient policies in discrete dynamic programming: linear programming including suboptimality tests and additional constraints. Math Prog 30:46–70
10. Howard R (1960) Dynamic programming and Markov processes. MIT Press, Cambridge, MA
11. Jeroslow RG (1972) An algorithm for discrete dynamic programming with interest rates near zero. Manag Sci Res Report No 300. Carnegie-Mellon University, Pittsburgh
12. Jeroslow RG (1973) Asymptotic linear programming. Oper Res 21:1128–1141
13. Kallenberg LCM (1983) Linear programming and finite Markovian control problems. Math Centre Tract No 148 (Amsterdam)
14. Miller BL, Veinott AF (1969) Discrete dynamic programming with a small interest rate. Ann Math Statist 40:366–370
15. Smallwood RD (1966) Optimum policy regions for Markov processes with discounting. Oper Res 14:658–669
16. Stoer J, Bulirsch R (1980) Introduction to numerical analysis. Springer, Berlin Heidelberg New York
17. Waerden van der BL (1953) Modern algebra, vols 1 and 2. Frederick Ungar, New York
18. Zoutendijk G (1976) Mathematical programming methods. North-Holland, Amsterdam