MATHEMATICS OF OPERATIONS RESEARCH
Vol. 13, No. 3, August 1988
Printed in U.S.A.

AVERAGE, SENSITIVE AND BLACKWELL OPTIMAL POLICIES IN DENUMERABLE MARKOV DECISION CHAINS WITH UNBOUNDED REWARDS*†

ROMMERT DEKKER‡ AND ARIE HORDIJK§

In this paper we consider a (discrete-time) Markov decision chain with a denumerable state space and compact action sets and we assume that for all states the rewards and transition probabilities depend continuously on the actions.

The first objective of this paper is to develop an analysis for average optimality without assuming a special Markov chain structure. In doing so, we present a set of conditions guaranteeing average optimality, which are automatically fulfilled in the finite state and action model.

The second objective is to study simultaneously average and discount optimality as Veinott (1969) did for the finite state and action model. We investigate the concepts of n-discount and Blackwell optimality in the denumerable state space, using a Laurent series expansion for the discounted rewards. Under the same condition as for average optimality, we establish solutions to the n-discount optimality equations for every n.

1. Introduction. In this paper we consider a (discrete-time) Markov decision chain with a denumerable state space and compact action sets and we assume that for all states the rewards and transition probabilities depend continuously on the actions.

The first objective of this paper is to develop an analysis for average optimality without assuming a special Markov chain structure. In doing so, we present a set of conditions guaranteeing average optimality, which are automatically fulfilled in the finite state and action model.

The second objective is to study simultaneously average and discount optimality as Veinott (1969) did for the finite state and action model. We investigate the concepts of n-discount and Blackwell optimality in the denumerable state space, using a Laurent series expansion for the discounted rewards. Under the same condition as for average optimality, we establish solutions to the n-discount optimality equations for every n.

Our analysis consists of several parts. After introducing the model in §2, we establish the main result of this paper in §3: the existence of solutions to the so-called Blackwell optimality equations, from which the existence of average optimal policies follow. We show that this is the consequence of three conditions, the first is the existence of a Laurent series expansion for the total discounted rewards, the second is the existence of a good candidate policy, which follows from the continuity in the policy of the terms of the Laurent series, and finally the third is the basic property of the policy improvement procedure. This property states that an improving decision rule yields a

*Received January 3, 1984; revised May 18, 1987.
AMS 1980 subject classification. Primary: 90C40.
IAOR 1973 subject classification. Main: Programming: Markov Decision
OR/MS Index 1978 subject classification. Primary: 119 Dynamic Programming/Markov/Infinite State.
Key words. Denumerable Markov decision chains, average optimality, sensitive optimality criteria, optimality equation, unbounded immediate rewards, Laurent series expansion, multi-chain model.
†This research was partially supported by the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).
‡Koninklijke/Shell Laboratorium Amsterdam
§University of Leiden.

0364-765X/88/1303/0395$01.25

better policy. We assume that this property also holds for lexicographic improvements in Laurent series.

In §4 we elaborate on our conditions using a weighted supremum norm on the set of vectors on the state space. Assuming boundedness in this norm of the immediate rewards and of matrices related to the transition matrices, we set up an operator-theoretical approach of Laurent series and Blackwell optimality. In this approach the policy improvement property corresponds to the positivity of an appropriate operator. The approach can be seen as a generalization of the contraction operator approach of discount optimality. We investigate continuity of the terms of the Laurent series and present also a single condition, viz. uniform geometric convergence, under which all our assumptions are satisfied.

In §5 we extend the results of the previous section for stationary policies to nonstationary ones. Average optimality is established in the class of all policies. As criterion is used the lim sup in the Abel sense. It was pointed out in Schäl (1987) that also optimality with respect to Cesaro lim sups is guaranteed. This section also deals with n-discount optimality.

In §6 we show the verification of our conditions in two types of models. In these examples we use the geometric convergence of the transition probabilities to the stationary matrix in an appropriate norm. In the first model we give an extension of the finite Markov decision chain. The second model is an optimal stopping problem. In a subsequent paper we will use recurrence conditions to apply the results of §3.

We conclude this section with an historical overview of denumerable state Markov decision chains with respect to average and sensitive optimality and compare our results with those of others. With respect to average optimality, most authors assume some recurrence condition.

The simplest form of recurrence condition assumes a state to be positive recurrent under all policies. We will call this the *strong unichain case*. Average optimality for this case was analyzed by many authors, of which we mention Taylor (1965), Derman (1966), Derman and Veinott (1967), Ross (1968), Hordijk (1974, 1976) and Tijms (1975).

The second form is the *unichain case*, in which there is one minimal closed set under all policies and for each policy there is a state to which there is appropriate recurrence. This model was treated by Federgruen, Hordijk and Tijms (1979a). In both cases the average rewards are the same for all states, which simplifies the analysis considerably.

Relaxing the unichainedness assumption implies allowing multiple classes. Although this complicates the analysis considerably, it does not in the *communicating case*, where for each pair of states, i, j, there exist policies f, g such that i is accessible from j under policy f and j from i under policy g (i and j are said to *communicate*). Since for average optimality only the rewards on the long term count, the short term rewards are irrelevant. This implies, that if an average optimal policy exists, an average optimal policy with unichain transition structure also exists by the communicatingness. The optimal average rewards are again the same for all states. This case was analyzed both by Bather (1973) for the finite state space and by Hordijk (1972, 1974) for the denumerable state space. If one assumes sufficient recurrence towards a finite set M, one only needs to assume that the states within M communicate, which is a weaker form of the communicating case. This case was treated by Federgruen, Schweitzer and Tijms (1983), who also established the existence of solutions to the average optimality equations. This assumption is natural in most queueing and inventory models.

We speak of the *multichain case* if there is no specific Markov chain structure. In addition, we distinguish one special type called the *strong multichain case* which is the case when there is a set of states which is always reached from any state and under any policy. States within this set are not accessible from each other under any policy. The strong unichain case is a special case of this model. Start-ups for an analysis of the multichain case were given by Hordijk (1974) and Wijngaard (1977). Recently this case got attention from several authors. Deppe (1985) considers an extension of the Markov decision chain model and shows the existence of an average optimal policy, but not of solutions to the average optimality equations. This is shown in Zijm (1985) under recurrence conditions for the bounded rewards case. Zijm (1984) gives an extension to the unbounded case. In Federgruen, Hordijk and Tijms (1978, 1979b) and Zijm (1985) equivalence results between a number of recurrence conditions, including uniform geometric convergence are given.

Kadota seems to have shown the existence of a stationary 0-discount optimal policy in the bounded rewards case under a quasi-compactness condition together with a uniform convergence condition. Dietz and Nollau (1983) investigate policy improvement under a uniform Doeblin condition, which yields characterizations of average and sensitive optimality, but no existence results. In Sennott (1986) average optimality for nonnegative costs is analyzed via the Taylor-Ross approach.

In Mann (1983) the existence of solutions to the n-discount optimality equations are shown in the bounded rewards case under the condition that the number of minimal closed sets is continuous in the policy and that the mean recurrence time to a finite set is bounded (these conditions were shown to be equivalent in the aperiodic case to uniform geometric convergence, cf. Zijm 1985). Mann (1984) claims (the paper does not contain proofs) to have conditions sufficient to establish n-discount optimality also in the unbounded rewards case. In Hordijk and Dekker (1983) it is shown that the simultaneous Doeblin condition alone is not sufficient for the existence of an average optimal policy. That paper is also a set-up for this paper. Neither Zijm nor Mann prove the optimality within the class of nonstationary policies. Their analysis starts from recurrence conditions and requires that a finite set is eventually reached regardless of the starting state. This paper is based on an operator theoretical approach because of its generality and elegance. A significant difference is that we do not require a special Markov chain structure, the number of minimal closed sets may be infinite, the chains may be dissipative and even total reward convergent Markov decision chains are included. An analysis based on recurrence conditions is far more technical. Apart from average optimality this paper mainly deals with Blackwell optimality. Average optimality does not require a complete Laurent series expansion but only the first two terms and recurrence conditions guaranteeing only that, are slightly more general. Average optimality is however a rather rough criterion; short-term rewards are not important and only the long-term rewards matter. This calls for more sensitive optimality criteria of which Blackwell optimality is a good candidate. However, for Blackwell optimality we need stronger conditions, e.g., the communicating case is not sufficient (cf. Bather 1973). Our approach with weighted supremum norms and positive operators has many similarities with the contraction operator approach for discount optimality cf. Blackwell (1965) and Denardo (1967). In fact, the discounted case can be considered as a special case of our analysis: for a fixed interest rate $\rho = \rho_1$, we replace all Laurent series in ρ and operators depending on ρ by their value in ρ_1. We will come back to this at the end of §4.

2. The model.

Consider a dynamic system which is observed by a decision maker at discrete time points $t = 1, 2, \ldots$ to be in one of the states of a denumerable *state space* E. If at time point t, the system is observed in state i, the decision maker controls the system by choosing an action from a set $A(i)$, the *set of available actions*, which is independent of t. If action a is chosen in state i, then the following happens, independently of the history of the process:

(i) a *reward* $r_i(a)$ is earned immediately.

(ii) the system will be in state j at the next time point with *transition probability* $P_{ij}(a)$ ($P_{ij}(a) \geq 0$, $j \in E$ and $\Sigma_j P_{ij}(a) \leq 1$).

Note that the transition probability distribution may be defective i.e., $\Sigma_j P_{ij}(a) < 1$. In this way our analysis includes the total reward model. The results from Markov (decision) chain theory which are used in this paper remain valid under this relaxation.

A *decision rule* π^t at time t is a function that assigns to each action the probability of that action being taken at time t; in general, it may depend on all realized states up to and including time t and all realized actions up to time t. A *policy* R is a sequence of decision rules $R = (\pi^1, \pi^2, \dots, \pi^t, \dots)$. For a *memoryless* or *Markov* policy the decision rule at time t is independent of the states and actions before time t. A policy is said to be *stationary* and *deterministic* if all decision rules are identical and nonrandomized. Hence a stationary and deterministic policy is completely described by a mapping f: $E \to \bigcup_{i \in E} A(i)$ such that $f(i) \in A(i)$ for each $i \in E$. These policies are denoted properly by $f^{(\infty)}$, however, we shall skip the (∞) symbol to keep the notation short.

Let C be the class of all policies and C_M the class of Markov policies. We claim that we prove our optimality results in the class of all policies. However, in notations and proofs we restrict ourselves to the class of Markov policies. There is no loss in generality in doing so. The reason is that the Derman-Strauch theorem (Derman 1970, Theorem 7.1) can be extended to the denumerable state space case (Hordijk 1974, Theorem 13.2). This means that to any policy and starting state there is a Markov policy having the same marginal distribution of the state and the action for any time point t. Consequently, for the optimality criteria we use in this paper optimality in C is equivalent to optimality in C_M. In §§3 and 4 we shall restrict ourselves to stationary and deterministic policies. Let F denote the class of these policies and observe that F can be represented by $\Pi_{i=1}^{\infty} A(i)$. With respect to the actions we assume

Assumption 1. (i) $A(i)$ is a compact metric set for all $i \in E$.

(ii) Both $r_i(a)$ and $P_{ij}(a)$ are continuous on $A(i)$ for all $i, j \in E$.

Observe that by Assumption 1(i) F is a compact set in the product topology. Before introducing optimality criteria we give further notation. Let e denote the vector (on E) with all components equal to one and I be the *identity matrix*. For any decision rule π we denote by $P(\pi)$, $r(\pi)$ the matrix, vector with $P_{ij}(\pi(i))$ and $r_i(\pi(i))$ respectively as elements. Let $P^k(R) = P(\pi^1) \cdot P(\pi^2) \cdots P(\pi^k)$ and $P^0(R) = I$. When a vector is denoted by x, then $(x)_i$ is its ith component. We call $A(f)$ and $x(f)$ a *matrix* and *vector function* if for any $f \in F$, $A(f)$ and $x(f)$ is a matrix and vector on E respectively. For vectors x, y we say $x \geq y$ if $x_i \geq y_i$, $i \in E$. We consider the following criteria for an infinite time horizon.

$$v^{\alpha}(R) := \sum_{k=0}^{\infty} \alpha^k P^k(R) r(\pi^{k+1}) \tag{2.1}$$

which is called the *expected α-discounted rewards* of policy R. The factor α is taken from $[0, 1)$ and is called the *discount factor*. In the sequel we shall also use $\rho := (1 - \alpha)/\alpha$ or equivalently $\alpha = 1/(1 + \rho)$, where ρ is called the *interest rate*. We also write $v^{\rho}(R)$ for the $1/(1 + \rho)$-discounted rewards, so the symbols α and ρ denote the discounting with a factor α and $1/(1 + \rho)$. A policy R_{α} is *α-discounted optimal* if for all $i \in E$, $R \in C$

$$v_i^{\alpha}(R_{\alpha}) \geq v_i^{\alpha}(R). \tag{2.2}$$

A second criterion is the *(long-run) average expected rewards*, defined by

$$(2.3)\qquad g(R) := \liminf_{N\to\infty} \frac{1}{N+1} \sum_{k=0}^{N} P^k(R) r(\pi^{k+1}).$$

So far, expressions (2.1) and (2.3) may be undefined; we shall give conditions guaranteeing its existence later. A policy R_0 is *average optimal* if for all $i \in E$, $R \in C$, $g_i(R_0) \geqslant g_i(R)$. It is well known that average optimality is a rather insensitive criterion, only the rewards in the long run are of importance. Several more sensitive criteria, i.e. criteria that imply average optimality, were suggested for the finite state and action model. We shall generalize some of these criteria to the denumerable state and compact action model. Veinott (1969) introduced *n-discount optimality*: a policy R_0 is n-discount optimal, $n \geqslant -1$, if for all $i \in E$, $R \in C$

$$\liminf_{\rho\downarrow 0} \rho^{-n}\left[v_i^\rho(R_0) - v_i^\rho(R)\right] \geqslant 0.$$

Blackwell (1962) introduced the concept of the 1-*discount-optimal policy*, which was later renamed *Blackwell optimal policy*. In the denumerable state and compact action case we discern two versions of this criterion, called *Blackwell* and *strong Blackwell optimality* respectively.

DEFINITION 2.1. (i) A policy R_0 is Blackwell optimal if for every $i \in E$, $R \in C$ there exists a $\rho(i, R) > 0$ such that $v_i^\rho(R_0) \geqslant v_i^\rho(R)$, $0 < \rho \leqslant \rho(i, R)$.

(ii) A policy R_0 is strongly Blackwell optimal if

$$\rho_0 := \inf_{i\in E, f\in F} \rho(i, f) > 0.$$

Strong Blackwell optimality corresponds to the version introduced by Blackwell (1962). In the finite state and action case both criteria are the same. For that case it can be shown that one can restrict oneselves to the class of stationary and deterministic policies with respect to α-discount or average optimality. The relation between the forementioned optimality criteria in that case can be made clear in the following way. First, it is shown that $v^\rho(f)$ is the unique solution to the equation

$$(2.4)\qquad v = r(f) + \frac{1}{1+\rho} P(f) v.$$

This equation can be rewritten as $[I - P(f)/(1+\rho)]v = r(f)$. Since $[I - P(f)/(1+\rho)]$ is a nonsingular matrix, we can invert it and obtain

$$(2.5)\qquad v^\rho(f) = \left[I - \frac{1}{1+\rho} P(f)\right]^{-1} r(f)$$

$$= (1+\rho)\left[(1+\rho)I - P(f)\right]^{-1} r(f).$$

Using Cramer's rule, we see that $v^\rho(f)$ can be written as a rational functional of ρ. This implies that we can write $v^\rho(f)$ as a Laurent series in ρ, given by

$$(2.6)\qquad v^\rho(f) = (1+\rho) \sum_{k=-1}^{\infty} \rho^k y^{(k)}(f)$$

(cf. Miller and Veinott 1969).

It follows from a Tauberian theorem, (cf. Denardo and Miller 1968) that $y^{(-1)}(f) = g(f)$, implying that -1 -discount optimality is equivalent to average optimality. From (2.6) we see that n-discount optimality corresponds to lexicographically maximizing the first $n+2$ terms of the Laurent series expansion for $v_i^\rho(f)$, for all $i \in E$. Blackwell optimality is the same as lexicographic optimality of all terms of the Laurent series for $v_i^\rho(f)$ for each $i \in E$, which implies n-discount optimality for all $n = -1, 0, \ldots$.

Under some assumptions this also holds in the denumerable space, as will be shown in §5. Moreover, it appears possible to define Blackwell optimality equations characterizing Blackwell optimal policies.

The above analysis does not carry over to the denumerable state model for two reasons. Firstly, we are dealing with infinite matrices which implies that (2.5) is no longer valid. Secondly, we want to allow $r_i(f)$ to be unbounded in $i \in E$; however, this immediately creates problems with respect to the finiteness of the terms of the Laurent series.

3. Concepts for Blackwell optimality.

In this section we provide a general framework not only for Blackwell but also for average and n-discount optimality. The analysis is independent of the Markov chain structure, allowing multiple minimal closed sets as well as a communicating case. This section only deals with the general concepts for the framework as to keep the approach as clear and broad as possible. A detailed analysis requires a number of technical assumptions which will be given in the following section.

The concepts from which we establish Blackwell optimality are firstly the existence of a Laurent series expansion for the total discounted rewards, secondly, the existence of a good candidate (mostly implied by continuity of the terms of the Laurent series) and thirdly, policy improvement, lexicographically for all terms of the Laurent series expansion.

We shall restrict ourselves in this chapter to the class of stationary policies, the extension to the class of nonstationary policies will be given in §5.

We start with the existence of a Laurent series expansion for the total discounted rewards of stationary policies. Although such an expansion always exists in the finite state and action model (cf. 2.6) it has to be proven or assumed in the denumerable state space model.

Condition 2. For each $f \in F$ there exists a $\rho(f) > 0$ such that

$$v^\rho(f) = (1+\rho)\sum_{k=-1}^{\infty} \rho^k y^{(k)}(f), \qquad 0 < \rho \leqslant \rho(f). \tag{3.1}$$

The analysis following is in many ways analogous to an analysis for the total discounted rewards, the main difference being that the last mentioned analysis treats $v^\rho(f)$ only for a fixed value of ρ and therefore as a real vector on E, while in our analysis $v^\rho(f)$ will be treated as a vector of Laurent series on E.

To that end we will introduce a set and a linear space of Laurent series on which we define operators similar to the well-known operators L_f, U from the discounted rewards analysis (cf. Blackwell 1967 and Denardo 1967). Let x be a Laurent series and $x(\rho)$ its value in ρ. Let

$$LS := \left\{ x = x(\rho) \;\middle|\; \begin{array}{l} x(\rho) = (1+\rho)\sum_{k=-1}^{\infty} \rho^k a^{(k)},\ a^{(k)} \in \mathbf{R} \\ \limsup_{k\to\infty} |a^{(k)}|^{1/k} < \infty \end{array} \right\} \quad \text{and}$$

$$LS^E := \{ y \mid y_i \in LS,\ i \in E \}$$

be a corresponding linear space of Laurent series.

By $v^\rho(f)$ we denote both the Laurent series as well as its value in ρ. Two elements of LS can be added by adding terms of the Laurent series with the same power of ρ. An element of LS is multiplied by a scalar by multiplying all terms of the Laurent series with that scalar. Note that a Laurent series of the form $\Sigma_{k=-1}^{\infty}\rho^k a^{(k)}$ can be rewritten into an element of LS through dividing by a factor $(1+\rho)$. This factor in the definition of LS is therefore not crucial but only used for convenience. Elements of LS and LS^E will be ordered in the following way.

DEFINITION 3.1. (a) $x\,{}_l\!\geqslant y$, $x, y \in LS$ if $\liminf_{\rho\downarrow 0}\rho^{-k}[x(\rho)-y(\rho)] \geqslant 0$, $k = -1, 0, \ldots,$

$x\,{}_l\!> y$, $x, y \in LS$ if $x\,{}_l\!\geqslant y$ and $x\,{}_l\!\neq y$.

(b) $x\,{}_l\!\geqslant y$, $x, y \in LS^E$ if $x_{i\,l}\!\geqslant y_i$, $i \in E$,

$x\,{}_l\!> y$, $x, y \in LS^E$ if $x_{i\,l}\!\geqslant y_i$, $i \in E$ and $x_{i_0\,l}\!> y_{i_0}$, for some $i_0 \in E$.

Note that $x\,{}_l\!> 0$, $x \in LS$ is equivalent to the first nonzero term of the Laurent series for x being positive (the factor $(1+\rho)$ is not of influence here). Remark further that $x\,{}_l\!\geqslant y$, $x, y \in LS$ is equivalent to $x(\rho) \geqslant y(\rho)$ for $0 < \rho \leqslant \rho_0$ for some $\rho_0 > 0$.

This implies that under Condition 2 a policy f_0 is Blackwell optimal within the class F if $v^\rho(f_0)\,{}_l\!\geqslant v^\rho(h)$ for all $h \in F$. Moreover, for each state $i \in E$ the Laurent series for $v_i^\rho(f_0)$ is maximal in the sense of lexicographic ordering.

Analogously to the discounted rewards case we want to make use of policy improvement, however, in this case, for Laurent series. Well known in the discounted rewards case is the following operator

$$L_f^\rho v = r(f) + \frac{P(f)}{1+\rho}v - v, \qquad v \in \mathbf{R}^E \tag{3.2}$$

for $f \in F$. We will define its generalisation for Laurent series, denoted by L_f:

$$L_f y\,{}_l\!= r(f) + \frac{P(f)}{1+\rho}y - y, \qquad y \in LS^E \tag{3.3}$$

or equivalently for ρ small,

$$\begin{aligned} L_f y(\rho)\,{}_l\!= {} & \rho^{-1}\{P(f)y^{(-1)} - y^{(-1)}\} \\ & + \rho^0\{r(f) + P(f)y^{(0)} - y^{(0)} - y^{(-1)}\} \\ & + \sum_{k=1}^{\infty}\rho^k\{P(f)y^{(k)} - y^{(k)} - y^{(k-1)}\}. \end{aligned}$$

Note that $P(f)y^{(k)}$ does not necessarily exist for all $y^{(k)} \in \mathbf{R}^E$; in that case L_f is restricted to an appropriate subset of LS^E. The operator L_f yields the one-step improvement in Laurent series under policy f. For the discounted rewards case it is well known that $L_f^\rho v > 0$ implies that $v^\rho(f) > v$, which is the basis for policy improvement. For the operator L_f on LS^E a similar property can only be proven once it is specified how the terms of the Laurent series relate to the transition matrix $P(f)$. This will be done in the next section, whereas we will assume it in this section.

Condition 3. If for some $f \in F$, $y \in LS^E$, $L_f y$ exists and $L_f y\,{}_l\!> 0$ then $v^\rho(f)\,{}_l\!> y$. The same implication also holds if the "${}_l\!>$" relation is replaced by either the "${}_l\!=$" or "$<_l$" relations.

In the finite state and action case this condition follows directly from policy improvement for discounted rewards with a fixed ρ. $L_f y\,{}_l\!> 0$, implies in that case $L_f^\rho y(\rho) > 0$ for $0 < \rho \leqslant \rho_0$ for some $\rho_0 > 0$. From policy improvement with discounted rewards it follows that $v^\rho(f) > y(\rho)$, $0 < \rho \leqslant \rho_0$, hence, $v^\rho(f)\,{}_l{>}\, y$. The other relations are proven similarly.

In the denumerable state space this argument no longer satisfies, since ρ_0 is obtained as a minimum over states, which can be zero in a denumerable state space.

Apart from the one-step improvement operator L_f working with a fixed policy f, we also introduce an operator U for the maximal one-step improvement, viz.

$$(3.4)\qquad Uy\,{}_l{=}\ \operatorname{lex}\cdot\sup_{f\in F} r(f) + \frac{P(f)}{1+\rho}y - y, \qquad y \in LS^E.$$

Lex · sup denotes lexicographic supremum, which is taken for each state separately. Inserting the Laurent series expression for y, it means that we determine

$$U_i^{(-1)}y := \sup\left\{\sum_j P_{ij}(a)y_j^{(-1)} - y_i^{(-1)} | a \in A(i)\right\}$$

and consecutively of

$$U_i^{(0)}y := \sup\left\{r_i(a) + \sum_j P_{ij}(a)y_j^{(0)} - y_i^{(0)} - y_i^{(-1)} | a \in A^{(-1)}(i)\right\} \quad \text{where}$$

$$A^{(-1)}(i) := \left\{a \in A(i) | \sum_j P_{ij}(a)y_j^{(-1)} = \sup_{\tilde{a}\in A(i)} \sum_j P_{ij}(\tilde{a})y_j^{(-1)}\right\}$$

and of

$$U_i^{(k)}y := \sup\left\{\sum_j P_{ij}(a)y_j^{(k)} - y_i^{(k)} - y_i^{(k-1)} | a \in A^{(k-1)}(i)\right\}$$

for $k = 1, 2, \dots$, where $A^{(k-1)}(i)$, $k = 0, 1, \dots$ is the subset of $A^{(k-2)}(i)$ consisting of the maximizing actions of the terms $r_i(a) + \Sigma_j P_{ij}(a)y_j^{(0)}$ for $k = 0$ and of $\Sigma_j P_{ij}(a)y_j^{(k)}$ for $k = 1, 2, \dots$. Uy is obtained by combining the terms $U^{(k)}$ into a Laurent series: $Uy\,{}_l{=}\, \Sigma_{k=-1}^{\infty}\rho^k U^{(k)}y$. Similar as for the operator L_f, U may be defined only on a subset of LS^E. Note that $Uv^\rho(f)\,{}_l{\geqslant}\, 0$, $f \in F$, since by Condition 3

$$(3.5)\qquad r(f) + \frac{P(f)}{1+\rho}v^\rho(f) - v^\rho(f)\,{}_l{=}\, 0.$$

This leads us to consider the following set of equations in LS^E.

$$(3.6)\qquad Uy\,{}_l{=}\, 0,$$

which for obvious reasons we call the *Blackwell optimality equations*.

The first result will be the existence of a solution $y \in LS^E$ to the Blackwell optimality equations, or even stronger, the existence of a policy f_0 for which $v^\rho(f_0)$ is a solution. For this result, it appears that Conditions 2, 3 alone are not sufficient. Using Condition 3 one could apply policy improvement starting from some initial policy. However, convergence of the sequence of policies is not guaranteed, let alone convergence to optimal values (cf. Examples 3.1.1 and 3.2.1 in Dekker 1985). In this section therefore, we will make an extra assumption: the existence of a good candidate. The way in which this condition will be used originates from an idea of Schweitzer (1982).

Condition 4. There exists a policy f_0 which maximizes lexicographically $y_i^{(k)}(f)$, $(i, k) \in E \times \{-1, 0, 1, \ldots\}$ in an enumeration of $E \times \{-1, 0, 1, \ldots\}$ such that (i, k) comes before (i, l) if $k \leqslant l$.

In Condition 4 we use an enumeration of the terms of the Laurent series for $v^\rho(f)$ over all states, in which for each state the terms of the Laurent series are enumerated with increasing powers.

Condition 4 is a consequence of the following condition.

Condition 4a. $y_i^{(k)}(f)$ is a continuous function in f for all $i \in E$, $k = -1, 0, 1 \ldots$.

THEOREM 3.1. *Condition* 4a *implies Condition* 4.

PROOF. Let $x(f)$ be a real continuous function of f. Since F is compact, there exists a policy f_1 maximizing $x(f)$ and the set $\{f \in F | x(f) = \max_{h \in F} x(h)\}$ is again compact. Condition 4a implies the existence of a sequence of nonincreasing compact subsets $F_1, F_2, \ldots$ in which policies in F_k, $k = 1, 2, \ldots$ maximize the first k terms of the enumeration in Condition 4 lexicographically. Remark further that none of these subsets F_k is empty. From standard analysis it follows that $F_\infty = \bigcap_{k=1}^{\infty} F_k$ is also nonempty. Any policy from F_∞ has the properties required by Condition 4. ∎

Continuity is not the only way to produce a good candidate. If for a term in the enumeration there is an unique maximizing policy then Condition 4 is also satisfied. In a communicating case, Condition 4 can be fulfilled without having continuity of all terms of the Laurent series. However, this last case will be treated in a subsequent paper. In the finite state and action model, with only finitely many policies, Condition 4 is trivially met. Condition 4 appears to be sufficient as the following theorem shows.

THEOREM 3.2. *Under Conditions* 2, 3 *and* 4, $v^\rho(f_0)$, *with* f_0 *the policy from Condition* 4, *constitutes a solution to the Blackwell optimality equations. Moreover*, f_0 *is Blackwell optimal.*

PROOF. Suppose $Uv^\rho(f_0) {}_l{>} 0$ and say $(Uv^\rho(f_0))_i {}_l{>} 0$. Recall the detailed definition of the operator U, as is given after (3.4). Suppose further that the nth supremum is the first nonzero and so a positive one. Hence there exist some action $a_0 \in A(i)$ for which the first $(n-1)$ terms are zero and the nth term is positive.

Let $\tilde{f}$ be defined through $\tilde{f}(i) = a_0$, $\tilde{f}(j) = f_0(j)$, $j \neq i$. Hence $(L_{\tilde{f}} v^\rho(f_0))_i {}_l{>} 0$ and $(L_{\tilde{f}} v^\rho(f_0))_j {}_l{=} 0$, $j \neq i$. According to Condition 3 this implies that $v^\rho(\tilde{f}) {}_l{>} v^\rho(f_0)$, which implies that the enumeration of Condition 4 is larger for $\tilde{f}$ than for f_0, which is a contradiction. Since $Uv^\rho(f_0) {}_l{\geqslant} 0$ by (3.5) we have established that $Uv^\rho(f_0) {}_l{=} 0$, and $v^\rho(f_0)$ is a solution. Accordingly we have for any $f \in F$: $L_f v^\rho(f_0) \leqslant_l 0$ and by Condition 3, $v^\rho(f) \leqslant_l v^\rho(f_0)$, $f \in F$, which implies that policy f_0 is Blackwell optimal. ∎

An even more important aspect of the Blackwell optimality equations is stated in the following theorem. Remark that it does not require Condition 4!

THEOREM 3.3. *Suppose Conditions* 2 *and* 3 *hold and let* y *be a solution to the Blackwell optimality equations. Any conserving policy* f, *i.e., policy* f *for which* $L_f y {}_l{=} 0$ *is Blackwell optimal and*

$$v^\rho(f) {}_l{=} y {}_l{\geqslant} v^\rho(h), \qquad h \in F. \tag{3.7}$$

PROOF. By Condition 3 $L_f y {}_l{=} 0$ implies that $v^\rho(f) {}_l{=} y$. For any other policy $h \in F$ obviously $L_h y \leqslant_l 0$, hence $v^\rho(h) \leqslant_l y {}_l{=} v^\rho(f)$. ∎

Note that Theorem 3.3 implies the uniqueness of the solution of the Blackwell optimality equations if there are conserving policies. An extension of Theorem 3.3 to nonstationary policies will be given in §5.

If we use the detailed expansion of the U operator, as is given after (3.4), in the Blackwell optimality equations, we observe that the first two equations constitute the so-called average optimality equations. The first $(n + 3)$ equations will constitute the n-discount optimality equations. Since these only require $n + 3$ terms of the Laurent series it suffices to assume only their existence instead of the entire Laurent series. Condition 3 can accordingly be restricted to only the first $n + 3$ terms of the Laurent series. We will come back to this in §5. Recall that in the finite state and action model Conditions 2, 3, and 4 are fulfilled, which justifies the introduction of Blackwell optimality equations in that model.

4. Blackwell optimality through normed linear spaces.

4.1. *Positive operators on Laurent series.* In this section we shall give a detailed analysis of Laurent series and conditions required for optimality. We start with introducing a weighted supremum norm $\|\cdot\|_\mu$, since we want to allow for unbounded (in the state) immediate rewards. It is defined by

$$\|x\|_\mu := \sup_{i \in E} \frac{|x_i|}{\mu_i}, \tag{4.1}$$

for any vector x on E, where $\mu_i > 0$, $i \in E$, are positive weights. Let $\|A\|_\mu$ be the associated operator norm. If A is a matrix on E, we can write

$$\|A\|_\mu = \sup_{i \in E} \frac{1}{\mu_i} \sum_j |A_{ij}| \mu_j .$$

Based on this norm we introduce normed linear spaces V^μ, M^μ of vectors, matrices on E respectively. These spaces guarantee that for any matrix $A \in M^\mu$ and vector $x \in V^\mu$, Ax exists and is again contained in V^μ. Our approach is to look for such a vector μ that not only $P(f)$ and $r(f)$, $f \in F$ have finite μ-norms, but that also each term of the Laurent series expansion for $v^\rho(f)$, $f \in F$ (cf. Condition 2) has a finite μ-norm. To that end we first investigate how the terms of the Laurent series expansion relate to the matrix of transition probabilities $P(f)$. Since the average rewards play an important role in the Laurent series expansion, we first consider the long-term behaviour of the Markov chain. This can be described by the matrix $\Pi(f)$, whose elements are defined as

$$\Pi_{ij}(f) := \lim_{N \to \infty} \frac{1}{N+1} \sum_{k=0}^{N} P_{ij}^k(f), \qquad i, j \in E, f \in F. \tag{4.2}$$

$\Pi(f)$ always exists in a denumerable state space (cf. Chung 1960). It is called the *stationary matrix* since the following relations hold:

$$P(f)\Pi(f) = \Pi(f)P(f) = \Pi(f)\Pi(f) = \Pi(f), \qquad f \in F. \tag{4.3}$$

At this point we have to make some remarks on Abelian and Cesaro limits. Given a sequence $a_k \in \mathbf{R}$, $k = 0, 1, 2, \ldots$, its *Abelian limit* A is defined by

$$A = \lim_{\alpha \uparrow 1} (1 - \alpha) \sum_{k=0}^{\infty} \alpha^k a_k,$$

its *Cesaro* limit C as

$$C = \lim_{N\to\infty} \frac{1}{N+1} \sum_{k=0}^{N} a_k,$$

provided the expressions exist.

The Abelian and Cesaro limits are extensions of the normal limits, i.e., if $a_k \to a$ in the normal sense, then both A and C exist and are equal to a. The Abelian limit is the weakest one, i.e., if the Cesaro limit exists then the Abelian too and they are equal. The converse is in general not true, however, if all $a_k \geqslant 0$, $k = 0, 1, \ldots$ then existence of the Abelian limit implies existence of the Cesaro limit and their equality (cf. Titchmarsh 1939).

The matrix $\Pi(f)$ is closely related to the average rewards, $g(f) := \liminf_{N\to\infty} 1/(N+1) \sum_{k=0}^{N} P(k) r(f)$, the lim inf taken component wise.

In this chapter we shall make assumptions ensuring that the lim inf in the definition can be replaced by a limit, and moreover, $g(f) = \Pi(f) r(f)$.

Apart from the average rewards we also consider n-discount and Blackwell optimality. Hence we have to take into account the rewards above or below the average rewards, caused by deviations of the Markov chain from the stationary behaviour. This is expressed by the matrix $D(f)$, whose elements are defined by

$$(4.4) \qquad D_{ij}(f) = \lim_{\alpha\uparrow 1} \sum_{k=0}^{\infty} \alpha^k \left[P_{ij}^k(f) - \Pi_{ij}(f)\right], \qquad i, j \in E, f \in F,$$

provided the limit exists.

$D(f)$ is defined through an Abel summation to incorporate periodicity of the Markov chain. It is called the *deviation matrix* (cf. Veinott 1969) or also the *ergodic potential operator* (cf. Syski 1978). In the finite state and action case $D(f)$ always exists and it can be used to give the Laurent series expansion for the discounted rewards (cf. Veinott 1974):

$$(4.5) \quad v^{\rho}(f) = (1+\rho)\left[\frac{\Pi(f) r(f)}{\rho} + \sum_{k=0}^{\infty} (-1)^k \rho^k D^{k+1}(f) r(f)\right], \qquad f \in F.$$

To guarantee that this expansion also holds in a denumerable state space, we have to make some assumptions. Firstly, that the discounted rewards $v^{\rho}(f)$ exists for ρ close to zero. Some authors have given sufficient conditions for the existence of $v^{\rho}(f)$ for a fixed ρ, but most of these conditions cannot be fulfilled for all ρ close to zero. Secondly we assume that the matrix $D(f)$ exists and that all necessary relations with $D(f)$ hold. Finally, we assume that criteria like $v^{\rho}(f)$ and $g(f)$ can be obtained by operators working on the immediate rewards. In this way we come to the following condition.

Condition 5. (i) $c_0(f) := \|r(f)\|_{\mu} < \infty, \quad f \in F.$

(ii) $c_1(f) := \sup_{0<\alpha<1} (1-\alpha) \|\sum_{k=0}^{\infty} \alpha^k P^k(f)\|_{\mu} < \infty$, $f \in F$.

(iii) The matrix $D(f)$ exists, $c_2(f) := \|D(f)\|_{\mu} < \infty$, $f \in F$ and the following relations hold

$$(4.6) \qquad [I - P(f)] D(f) = D(f)[I - P(f)] = I - \Pi(f).$$

$$(4.7) \qquad \Pi(f) D(f) = D(f) \Pi(f) = 0.$$

Note that $\sup_{k=1,2,\ldots} \|P^k(f)\|_{\mu} < \infty$, $f \in F$ implies finiteness of $c_1(f)$. In particular if $\mu = e$, then $\|P^k(f)\|_e \leqslant 1$, $k = 1, 2, \ldots$ and Condition 5(ii) is trivially fulfilled.

This can be used if $r_i(f)$ is bounded in $i \in E$. Condition 5(ii) can also be replaced by

$$\tilde{c}_1(f) := \sup_{N=1,2,\ldots} \left\| \frac{1}{N+1} \sum_{k=0}^{N} P^k(f) \right\|_\mu < \infty, \qquad f \in F,$$

but not by $\tilde{\tilde{c}}_1(f) := \|\Pi(f)\|_\mu < \infty$ (this may be true in a dissipative chain, while Condition 5(ii) is not fulfilled). Condition 5 enables us to define $v^\rho(f)$ for every $\rho > 0$, to give a characterization and to establish the Laurent series expansion (4.5).

THEOREM 4.1. *Suppose Condition 5 holds and let $\rho > 0$ and $f \in F$ be fixed. Then*
(i) *the power series for $v^\alpha(f)$ converges absolutely and*

$$\|v^\alpha(f)\|_\mu \leqslant \frac{1}{1-\alpha} c_0(f) c_1(f), \qquad f \in F. \tag{4.8}$$

(ii) *$v^\alpha(f)$ is the unique solution in V^μ to the set of equations*

$$v = r(f) + \alpha P(f) v. \tag{4.9}$$

(iii) *Moreover, the Laurent series expansion given in* (4.5) *is valid for* $0 < \rho < (c_2(f))^{-1}$.

PROOF. To show (i) we first consider $\Sigma_{k=0}^{\infty} \alpha^k \|P^k(f)\|_\mu \|r(f)\|_\mu$.
From Condition 5(ii) it follows that for any $\tilde{\alpha}$ with $\alpha < \tilde{\alpha} < 1$ we have

$$(1-\tilde{\alpha})\tilde{\alpha}^k \|P^k(f)\|_\mu \leqslant c_1(f), \qquad k = 0,1,2,\ldots. \tag{4.10}$$

Inserting this, together with $\|r(f)\|_\mu \leqslant c_0(f)$, yields

$$\sum_{k=0}^{\infty} \alpha^k \|P^k(f)\|_\mu \|r(f)\|_\mu \leqslant \sum_{k=0}^{\infty} \frac{\alpha^k}{\tilde{\alpha}^k} \frac{1}{(1-\tilde{\alpha})} c_1(f) c_0(f)$$

which is finite, since $\alpha < \tilde{\alpha}$.

This implies that the powerseries for $v^\alpha(f)$ converges absolutely and that we may write

$$\|v^\alpha(f)\|_\mu \leqslant \left\| \sum_{k=0}^{\infty} \alpha^k P^k(f) \right\|_\mu \|r(f)\|_\mu \leqslant \frac{c_1(f)}{1-\alpha} c_0(f).$$

The absolute convergence of the powerseries for $v^\alpha(f)$ directly implies that $v^\alpha(f)$ is a solution to (4.9). To prove the uniqueness, suppose that both $v, w \in V^\mu$ are solutions to (4.9). Subtraction of the equations yields $v - w = \alpha P(f)(v - w)$.

Iterating this k times and taking μ-norms gives $\|v - w\|_\mu \leqslant \alpha^{k+1} \|P^{k+1}(f)\|_\mu \|v - w\|_\mu$ for any $k \in \mathbb{N}$.

However, using (4.10) again for an $\alpha < \tilde{\alpha} < 1$ yields

$$\alpha^{k+1} \|P^{k+1}(f)\|_\mu \leqslant \frac{\alpha^{k+1}}{\tilde{\alpha}^{k+1}} \frac{c_1(f)}{(1-\tilde{\alpha})} \to 0 \qquad (k \to \infty),$$

which implies that $\|v - w\|_\mu = 0$.

For part (iii) we have to show that the right-hand side of (4.5) is a solution to (4.9). Using equations (4.3), (4.6) and (4.7) this can be proven with simple algebra, which we leave to the reader. ∎

A condition sufficient for Condition 5 will be presented in the following section. The purpose of Condition 5 is to allow that $v^\rho(f)$ is produced by an operator working on $r(f)$. To that end we first define a subspace LS_μ^E of LS^E, which is compatible for the μ-norm

$$LS_\mu^E := \Bigl\{(1+\rho)\sum_{k=-1}^{\infty}\rho^k y^{(k)} \bigm| y^{(k)} \in V^\mu,\ k=-1,0,\dots \text{ and}$$

$$\limsup_{k\to\infty}\left(\|y^{(k)}\|_\mu\right)^{1/k} < \infty\Bigr\}.$$

Consider the following operator

$$H^\rho(f) \;{}_l{:=}\; \Pi(f) + \sum_{k_l=0}^{\infty}(-1)^k\rho^{k+1}D^{k+1}(f). \tag{4.11}$$

Hence for $y \in LS_\mu^E$, $y(\rho) \;{}_l{=}\; (1+\rho)\Sigma_{k=-1}^{\infty}\rho^k y^{(k)}$

$$H^\rho(f)y(\rho) \;{}_l{=}\; (1+\rho)\Biggl\{\frac{\Pi(f)y^{(-1)}}{\rho} + \sum_{m=0}^{\infty}\rho^m\Bigl[\Pi(f)y^{(m)} + \sum_{k=0}^{m}(-1)^k D^{k+1}(f)y^{(m-k-1)}\Bigr]\Biggr\} \tag{4.12}$$

which is again a Laurent series with μ-bounded terms and contained in LS_μ^E (the check for the positive convergence radius is left to the reader).

$(1+\rho)H^\rho(f)/\rho$ can be considered as an inverse of $[I - P(f)/(1+\rho)]$ since it is easily verified that

$$\frac{(1+\rho)}{\rho}H^\rho(f)\Bigl[I - \frac{1}{1+\rho}P(f)\Bigr] \;{}_l{=}\; \Bigl[I - \frac{1}{1+\rho}P(f)\Bigr]\frac{(1+\rho)}{\rho}H^\rho(f) \;{}_l{=}\; I, \qquad f \in F. \tag{4.13}$$

The factor $(1+\rho)/\rho$ is not included in the definition to guarantee that $H^\rho(f)$ is an operator on LS_μ^E. This operator generates the Laurent series expansion for $v^\rho(f)$, i.e.

$$v^\rho(f) \;{}_l{=}\; \frac{(1+\rho)}{\rho}H^\rho(f)r(f), \qquad f \in F. \tag{4.14}$$

Remark that by Condition 5(ii) $\|P(f)\|_\mu \leqslant c_1(f)$, implying that $P(f)y$ exists for all $y \in LS_\mu^E$ and that accordingly L_f is a properly defined operator on LS_μ^E. By (4.13) and (4.14) we have

$$\frac{(1+\rho)}{\rho}H^\rho(f)L_f y \;{}_l{=}\; v^\rho(f) - y.$$

It now appears that Condition 3 is fulfilled if we can show that $H^\rho(f)$ is a positive operator for every $f \in F$, i.e., for all $y \in LS_\mu^E$

if $y \;{}_l{\geqslant}\; 0$ then $H^\rho(f)y \;{}_l{\geqslant}\; 0$ and

if $y \;{}_l{>}\; 0$ then $H^\rho(f)y \;{}_l{>}\; 0$, since $L_f y_l > 0$ then implies that $((1+\rho)/\rho)H^\rho(f)L_f y_l = v^\rho(f) - y_l > 0$ (the other relations follow in a similar way).

The proof that $H^\rho(f)$ is a positive operator is based on probabilistic arguments, in which we use the detailed expression (4.12). Note that for $x \in V^\mu$, $x > 0$ it is immediate that $H^\rho(f)x > 0$, since $(1+\rho)H^\rho(f)x/\rho$ are the discounted rewards in case x is the immediate reward vector, which are positive for all $\rho > 0$.

THEOREM 4.2. *$H^\rho(f)$ is a positive operator for each $f \in F$.*

PROOF. Suppose $y(\rho)_l = (1+\rho)\Sigma_{k=-1}^{\infty}\rho^k y^{(k)}{}_l \geqslant 0$. Let $i \in E$ and $f \in F$ be fixed and consider $(H^\rho(f)y)_i$. Let $B(i)$ denote the set of states accessible from i, i.e. $B(i) := \{j \in E | P_{ij}^n(f) > 0 \text{ for some } n \geqslant 0\}$. We divide $B(i)$ into two sets. Let $R(i) := \{j \in B(i) | \Pi_{ij}(f) > 0\}$ and $T(i) := \{j \in B(i) | \Pi_{ij}(f) = 0\}$. Note that it follows from its definition that $D_{ij}(f) > 0$ for $j \in T(i)$. Moreover, we have that $\Pi_{ij}(f) = 0$, $D_{ij}(f) = 0, \ldots,$ $D_{ij}^k(f) = 0, \ldots$ for $j \notin B(i)$, which implies that for $(H^\rho(f)y)_i$ we only have to consider y_j for $j \in B(i)$. For $y_i = (1+\rho)\Sigma_{k=-1}^{\infty}\rho^k y_j^{(k)}$, let $n(j)$ be the index of the first nonzero term, i.e., $y_j^{(m)} = 0$ for $-1 \leqslant m < n(j)$ and $y_j^{(n(j))} > 0$ (since $y_{jl} \geqslant 0$). Note that $n(j)$ may be infinite. Let $n_1 := \min\{\{n(j), j \in R(i)\}, \{n(j)+1, j \in T(i)\}\}$ and let $\bar{R}(i)$ and $\bar{T}(i)$ denote the subsets of $R(i)$ and $T(i)$ respectively in which the minimum is attained. If $n_1 = \infty$, then $y_{jl} = 0$ for all $j \in B(i)$ and according $(H^\rho(f)y)_{il} = 0$. Else, recall that (cf. (4.12))

$$(H^\rho(f)y)_{il} = (1+\rho)\Bigg\{\frac{1}{\rho}\sum_j \Pi_{ij}(f)y_j^{(-1)} + \sum_{m=0}^{\infty}\rho^m\Bigg[\sum_j \Pi_{ij}(f)y_j^{(m)}$$

$$+ \sum_{k=0}^{m}(-1)^k\sum_j D_{ij}^{k+1}(f)y_j^{(m-k-1)}\Bigg]\Bigg\}.$$

Since $y_j^{(m)} = 0$, $-1 \leqslant m < n_{-1}$ for $j \in R(i)$ and $y_j^{(m)} = 0$, $-1 \leqslant m < n_{-1}$ for $j \in T(i)$, it is easily verified that the first nonzero term of $(H^\rho(f)y)_i$ is given by

$$\rho^{n_1}\Bigg[\sum_{j \in \bar{R}(i)} \Pi_{ij}(f)y_j^{(n_1)} + \sum_{j \in \bar{T}(i)} D_{ij}(f)y_j^{(n_1-1)}\Bigg]$$

and that its coefficient is positive. Accordingly $(H^\rho(f)y)_{il} > 0$, which completes the first part of the proof. Remains to show that $y_l > 0$ implies that $(H^\rho(f)y)_j > 0$. If $y_l > 0$, there exists a state i for which $y_{il} > 0$. From the preceding we see that this implies that $(H^\rho(f)y)_{jl} > 0$ for every j with $i \in B(j)$. Since it is always true that $i \in B(i)$ (remark that $P_{ii}^0(f) = 1$) we have $(H^\rho(f)y)_{il} > 0$, which completes the proof. ∎

From the remarks preceding this theorem and from Theorem 4.1 it is clear that we have established

COROLLARY 4.3. *Condition 5 implies Conditions 2 and 3.*

4.2. *Uniform geometric convergence.* Although Condition 5 is a minimal assumption for Theorem 4.1 in the framework of a normed linear space, one wonders whether its verification can be facilitated. If $\mu = e$ (in the bounded rewards case) Condition 5 follows from a quasi-compactness condition for all policies which was studied in Wijngaard (1977) and Dietz and Nollau (1983). Closely related to, but stronger than, quasi-compactness is a uniform geometric convergence condition, denoted by μ-UGC (c, β), which however will yield more results.

Condition μ-UGC(c,β): There exist $c > 0$, $\beta < 1$ such that:

$$\begin{cases} \|P^k(f) - \Pi(f)\|_\mu \leqslant c\beta^k, & k = 1,2,\dots,\ f \in F \\ \|P(f)\|_\mu \leqslant c, & f \in F. \end{cases}$$

Note that this condition requires that on each minimal closed set under $P(f)$, the Markov chain is ergodic and aperiodic. Since the matrix $D(f)$ is defined through an Abelian limit, Condition 5 does not require aperiodicity. The μ-UGC condition can be altered to include periodic chains by introducing an initial distribution, but we will not elaborate this here. The condition does not restrict $\nu(f)$, the number of minimal closed sets under policy f. $\nu(f)$ may even be infinite.

Uniform geometric convergence in the normal supremum norm, denoted by e-UGC, was studied in Zijm (1985). In case $\Pi(f)$ is stochastic and $\nu(f) < \infty$ this condition was shown to be equivalent to a combination of the simultaneous Doeblin condition (cf. Hordijk 1974), aperiodicity and continuity of $\nu(f)$.

With respect to our analysis it appears that μ-uniform geometric convergence implies Condition 5(ii) and 5(iii) as is shown in the following theorem.

THEOREM 4.4. *The condition* μ-UGC(c,β) *implies the following assertions for every* $f \in F$.

(i) $\|\Pi(f)\|_\mu \leqslant 2c$,

(ii) $\sup_{0<\alpha<1}\|(1-\alpha)\sum_{k=0}^{\infty}\alpha^k P^k(f)\|_\mu \leqslant 2c + c/(1-\beta)$,

(iii) *The matrix* $D(f)$ *exists and* $D(f) = \sum_{k=0}^{\infty}[P^k(f) - \Pi(f)]$, *whereby the convergence is uniform in* f.

(iv) $\|D(f)\|_\mu \leqslant c/(1-\beta)$ *and relations* (4.6) *and* (4.7) *hold.*

PROOF. Assertion (i) is immediate since

$$\|\Pi(f)\|_\mu \leqslant \|P(f)\|_\mu + \|P(f) - \Pi(f)\|_\mu \leqslant 2c.$$

To prove assertion (ii) we remark that for fixed $0 < \alpha < 1$

$$(1-\alpha)\sum_{k=0}^{\infty}\alpha^k P^k(f) = (1-\alpha)\sum_{k=0}^{\infty}\alpha^k[P^k(f) - \Pi(f)] + \Pi(f), \qquad f \in F.$$

Hence,

$$\left\|(1-\alpha)\sum_{k=0}^{\infty}\alpha^k P^k(f)\right\| \leqslant \frac{(1-\alpha)c}{1-\alpha\beta} + 2c \leqslant \frac{c}{1-\beta} + 2c, \qquad f \in F.$$

With respect to assertion (iii) we first note that the series $\sum_{k=0}^{\infty}[P^k(f) - \Pi(f)]$ converges in μ-norm, uniformly in f, since by uniform geometric convergence it holds that

$$\left\|\sum_{k=N+1}^{\infty}[P^k(f) - \Pi(f)]\right\|_\mu \leqslant \frac{c\beta^{N+1}}{1-\beta}, \qquad f \in F.$$

It follows from an Abelian theorem that $D_{ij}(f)$ exists and is equal to $\sum_{k=0}^{\infty}[P_{ij}^k(f) - \Pi_{ij}(f)]$ which completes the proof of part (iii). From (iii) it directly follows that $\|D(f)\|_\mu \leqslant c/(1-\beta)$. The relations (4.6) and (4.7) follow in a standard way from the convergence in μ-norm of $\Sigma_{k=0}^{\infty}[P^k(f) - \Pi(f)]$. For example,

$$\|\Pi(f)D(f)\|_\mu = \left\| \Pi(f) \left\{ \sum_{k=0}^{N} [P^k(f) - \Pi(f)] + \sum_{k=N+1}^{\infty} [P^k(f) - \Pi(f)] \right\} \right\|_\mu$$

$$= \left\| \Pi(f) \sum_{k=N+1}^{\infty} [P^k(f) - \Pi(f)] \right\|_\mu \leqslant \frac{2c^2\beta^{N+1}}{1-\beta}, \qquad f \in F$$

for every $N \in \mathbb{N}$. Hence $\Pi(f)D(f) = 0$. ■

4.3. *Continuity.* One of the ways to verify Condition 4 is to establish continuity of the terms of the Laurent series expansion in the policy (cf. Theorem 3.1). Recall that the policy space F is defined as $\Pi_{i=1}^{\infty}A(i)$, the product space of the sets of available actions in each state and that F is endowed with the product topology. Each $A(i)$ is a compact metric set. Hence stationary policies converge, say $f^{(n)} \to f^{(0)}$ if $f^{(n)}(i) \to f^{(0)}(i)$ on $A(i)$, $i \in E$. If E is finite and $A(i)$ is finite for all i then F is also a finite set and any function on F is trivially continuous.

In our analysis we deal with vector and matrix functions in the policy and we shall use different kinds of continuity for them. A vector function $x(f)$ or a matrix function $A(f)$ is called pointwise continuous on E if $x_i(f)$ or $A_{ij}(f)$ is continuous for all $i, j \in E$. However, to establish continuity of a matrix vector product $A(f)x(f)$ it is not sufficient to have pointwise continuity of both $A(f)$ and $x(f)$. To that end we introduce the concept μ-continuity, which is related to the vector norm $\|\cdot\|_\mu$ used.

DEFINITION 4.1. A matrix function $A(f) \in M^\mu$ is μ-continuous on F if for every $i \in E$ and sequence $f^{(n)} \to f^{(0)}$ in F we have

$$\lim_{f^{(n)} \to f^{(0)}} \sum_j |A_{ij}(f^{(n)}) - A_{ij}(f^{(0)})| \mu_j = 0.$$

The following lemmas indicate the usefulness of μ-continuity and can be derived by standard analysis. Detailed proofs are given in Dekker (1985).

LEMMA 4.5. *For any matrix function $A(f) \in M^\mu$ the assertions* (i), (ii) *and* (iii) *are equivalent.*

(i) *$A(f)$ is μ-continuous on F.*

(ii) *$A(f)$ and $|A(f)|\mu$ are pointwise continuous on F.*

(iii) *For any sequence $x^{(n)} \to x^{(0)}$, pointwise converging on E with $\sup_{n=0,1,2,\dots} \|x^{(n)}\|_\mu < \infty$ it holds that $A(f^{(n)})x^{(n)} \to A(f^{(0)})x^{(0)}$ pointwise on E for every sequence $f^{(n)} \to f^{(0)}$ on F.*

LEMMA 4.6. *If both $A(f)$ and $B(f) \in M^\mu$ are μ-continuous on F, then*

(i) *$A(f) + B(f)$ is also μ-continuous*

(ii) *If $\sup_{f \in F}\|B(f)\|_\mu < \infty$, then $A(f)B(f)$ is also μ-continuous.*

Using the above lemma we can easily formulate a condition sufficient for condition 4a.

Condition 6. (i) *$P(f)\mu$ is pointwise continuous on F,*

(ii) *$D(f)$ is μ-continuous on F,*

(iii) *there exist constants c_0 and $c_2 \in \mathbb{R}$ with $\|r(f)\|_\mu \leqslant c_0$ and $\|D(f)\|_\mu \leqslant c_2$, $f \in F$.*

We now have

THEOREM 4.7. *Conditions* 5 *and* 6 *imply conditions* 2, 3 *and* 4a *and are sufficient to establish Blackwell optimality equations.*

PROOF. Since $P(f)$ is a nonnegative matrix, Conditions 5 and 6(i) imply by Lemma 4.5 that $P(f)$ is μ-continuous. From equations (4.6) and Lemma 4.6 we see that $\Pi(f)$ is also μ-continuous. Conditions 5 and 6 and Lemma 4.6 also imply that $D^k(f)$ is μ-continuous for $k = 2, 3, \ldots$. From Theorem 4.1 and Lemma 4.5 it now follows that every term $y^{(k)}(f)$, $k = -1, 0, \ldots$ of the Laurent series expansion for $v^\rho(f)$ given in (4.5), is pointwise continuous (Condition 4a). The final results follow from combining Corollary 4.3. Theorems 3.1, 3.2 and 3.3. ■

Remark that for any $y \in LS_\mu^E$, $y_i = (1+\rho)\Sigma_{k=-1}^{\infty}\rho^k y^{(k)}$, $P(f)y^{(k)}$ is pointwise continuous for all $k = -1, 0, \ldots$. Hence all suprema in the definition of the operator U, cf. (3.4), exist and can be replaced by maxima.

In the previous section we showed that Conditions 5(i) and 5(ii) were implied by uniform geometric convergence; however, the uniformness of the geometric convergence also has its consequences with respect to continuity.

THEOREM 4.8. *Conditions 5 and 6, sufficient to establish Blackwell optimality equations, are implied by the following set of conditions*

$$\begin{cases} \mu - UGC \\ \|r(f)\|_\mu \leqslant c_0, \qquad f \in F, \\ P(f)\mu \quad \text{pointwise continuous on } F. \end{cases}$$

PROOF. Note that by Theorem 4.7 we only have to prove that $D(f)$ is μ-continuous on F. To that end, we first prove that $\Pi(f)$ is μ-continuous on F. From the pointwise continuity of $P(f)\mu$ and $P(f)$ (Assumption 1) it follows by Lemma 4.5 that $P(f)$ is μ-continuous. Lemma 4.6 now implies that $P^k(f)$ is also μ-continuous for $k = 1, 2, \ldots$. Since condition μ-UGC (c, β) holds for some c and β, we have

$$\sum_j |P_{ij}^k(f) - \Pi_{ij}(f)|\mu_j \leqslant c\beta^k\mu_i, \qquad f \in F$$

which implies that both $P_{ij}^k(f)$ and $\Sigma_j P_{ij}^k(f)\mu_j$ converge uniformly in f to $\Pi_{ij}(f)$ and $\Sigma_j\Pi_{ij}(f)\mu_j$. Accordingly both $\Pi_{ij}(f)$ and $\Sigma_j\Pi_{ij}(f)\mu_j$, $i, j \in E$ are continuous in f. Hence $\Pi(f)$ is μ-continuous.

The μ-continuity of $D(f)$ can be established in a similar way from the uniform convergence of $\Sigma_{k=0}^N[P^k(f) - \Pi(f)]$ to $D(f)$, cf. Theorem 4.4. ■

REMARK 1. Although Theorem 4.1 establishes a Laurent series expansion for $v^\rho(f)$, it has not yet been shown that the first term of the expansion $\Pi(f)r(f)$ equals the average rewards $g(f)$.

Let

$$S^{(N)}(f) := \frac{1}{N+1}\sum_{k=0}^{N} P^k(f),$$

hence we have to show $S^{(N)}(f)r(f)$ converges pointwise to $\Pi(f)r(f)$ as $N \to \infty$. Recall that by its definition $S^{(N)}(f) \to \Pi(f)$ pointwise and that $\|r(f)\|_\mu \leqslant c_0(f)$ by Condition 5(i).

Analogously to Lemma 4.5 it can be proven that for pointwise convergence of $S^{(N)}(f)r(f)$ to $\Pi(f)r(f)$ we only need to establish pointwise convergence of $|S^{(N)}(f)|_\mu$ to $\Pi(f)_\mu$. Since $S_{ij}^{(N)}(f) \geqslant 0$, $i, j \in E$ and $\mu_j > 0$, $j \in E$, convergence in Cesaro sense of $S^{(N)}(f)\mu$ to $\Pi(f)\mu$ can be proven by the Abelian limit (cf. Titchmarsh 1939):

$$\lim_{\alpha \uparrow 1}(1-\alpha)\sum_{k=0}^{\infty}\alpha^k P^k(f)\mu = \Pi(f)\mu.$$

This follows directly from the operator $H^{\rho}(f)$, since,

$$(4.15)\qquad \sum_{k=0}^{\infty} \alpha^k P^k(f)\,\mu_{\,i} = \frac{(1+\rho)}{\rho} H^{\rho}(f)\mu$$

$$_{i}= (1+\rho)\left[\frac{\Pi(f)\mu}{\rho} + \sum_{k=0}^{\infty} (-1)^k \rho^k D^{k+1}(f)\mu\right].$$

The first equation from (4.15) is obtained in essentially the same way as the Laurent series expansion for $v^{\rho}(f)$, whereby $r(f)$ is replaced by μ.

In a similar way it can be proven that with respect to the *biasvector* $v(f)$, defined by

$$(4.16)\qquad v(f) := \lim_{\alpha\uparrow 1} \sum_{k=0}^{\infty} \alpha^k\left[P^k(f)r(f) - g(f)\right], \qquad f\in F,$$

we have

$$(4.17)\qquad v(f) = D(f)r(f), \qquad f\in F.$$

REMARK 2. The analysis in §§3 and 4 can be considered as a generalisation of the operator-theoretic approach for ρ-discount optimality whereby the contraction property is replaced by positivity. This can be illustrated in the following way. Conditions 5 and 6 are sufficient conditions for the existence and pointwise continuity of $v^{\rho}(f)$ for all $\rho > 0$. Consider a fixed interest rate ρ and replace all Laurent series and operators working on them by their value in ρ. The set LS^E is replaced by V^{μ}. Positivity of the $H^{\rho}(f)$ operator for a specific ρ is trivial and ρ-discount optimality equations can be established in the same way as the Blackwell optimality equations in Theorems 3.2 and 3.3.

5. Blackwell optimality within the class of nonstationary policies. In §§3 and 4 we only considered stationary policies and established optimality results within their class F. In this section we shall extend these results to the class C, including nonstationary policies. As explained in the introduction of §2 we may without loss of generality restrict the set of policies to C_M, the Markov policies. We shall assume that Conditions 5, 6 from §4 hold. To extend the implications of these conditions to the class of nonstationary policies we use a result which follows from the existence of nearly optimal stationary policies in positive dynamic programming (cf. Hordijk 1974, Theorem 13.6).

LEMMA 5.1. $\sup_{R\in C}\sum_{t=0}^{\infty}\alpha^t P^t(R)\mu = \sup_{f\in F}\sum_{t=0}^{\infty}\alpha^t P^t(f)\mu.$

This lemma implies the following extension of Condition 5(ii).

LEMMA 5.2. $\sup_{0<\alpha<1}(1-\alpha)\|\sum_{t=0}^{\infty}\alpha^t P^t(R)\|_{\mu} \leqslant c_1,\ R\in C_M.$

Similar to Theorem 4.1 we can prove that the power series for $v^{\alpha}(R)$ converges absolutely and that $\|v^{\alpha}(R)\|_{\mu} \leqslant c_0 c_1/(1-\alpha)$. However, $v^{\alpha}(R)$ has in general no Laurent series expansion. Using an approach developed by Hordijk and Sladky (1977) we shall expand $v^{\alpha}(R) - \hat{y}$ into the form $(1+\rho)\sum_{n=-1}^{\infty}\rho^n\phi^{(n)}(\alpha, R)$ with $\limsup_{\alpha\uparrow 1}(1-\alpha)^2\phi^{(n)}(\alpha, R) = 0$, for $n = -1, 0, \ldots$ and where $\hat{y}$ is the solution to the Blackwell optimality equations. This expansion enables us to prove that $\limsup_{\rho\downarrow 0}\rho^{-n}[v_i^{\rho}(R) - \hat{y}_i] \leqslant 0$, $i\in E$, $R\in C_M$, $n = -1, 0, \ldots$.

As a first step we introduce for any decision rule π vectors $\psi^{(-1)}(\pi), \psi^{(0)}(\pi), \ldots$ defined by (let $\hat{y}_l = (1+\rho)\Sigma_{n=-1}^{\infty}\rho^n\hat{y}^{(n)}$)

$$(5.1)\qquad \begin{cases} \psi^{(-1)}(\pi) := P(\pi)\hat{y}^{(-1)} - \hat{y}^{(-1)} \\ \psi^{(0)}(\pi) := r(\pi) + P(\pi)\hat{y}^{(0)} - \hat{y}^{(0)} - \hat{y}^{(-1)} \\ \psi^{(n)}(\pi) := P(\pi)\hat{y}^{(n)} - \hat{y}^{(n)} - \hat{y}^{(n-1)}, \qquad n = 1,2,\ldots. \end{cases}$$

Recall that by Theorems 3.2 and 4.7 there exists a Blackwell optimal stationary policy f_0 and that $v^{\rho}(f_0)_l = \hat{y}$. Moreover,

$$(5.2)\qquad \|\hat{y}^{(n)}\|_{\mu} \leqslant c(c_2)^n, \qquad n = -1,0,\ldots$$

for some constant c (recall that $\|D(f_0)\|_{\mu} \leqslant c_2$). Hence for any decision rule π and some constant $\tilde{c}$

$$(5.3)\qquad \|\psi^{(n)}(\pi)\|_{\mu} \leqslant \tilde{c}(c_2)^n, \qquad n = -1,0,\ldots.$$

Similarly to Theorem 4.1 it can be proven that $\Sigma_{t=0}^{\infty}\alpha^{t+1}P^t(R)\psi^{(n)}(\pi^{t+1})$ (with $R = (\pi^1, \pi^2, \ldots, \pi^{t+1}, \ldots)$) converges absolutely for all $n = -1,0,\ldots$. Furthermore, we have by Lemma 5.2

$$(5.4)\qquad \left\|\sum_{t=0}^{\infty}\alpha^{t+1}P^t(R)\psi^{(n)}(\pi^{t+1})\right\|_{\mu} \leqslant \frac{\alpha c_1}{1-\alpha}\tilde{c}(c_2)^n,$$

$$R \in C_M, \quad n = -1,0,\ldots$$

implying that

$$(5.5)\qquad \lim_{t\to\infty}\alpha^{t+1}P^t(R)\psi^{(n)}(\pi^{t+1}) = 0, \qquad R \in C_M, \quad n = -1,0,\ldots.$$

Moreover,

$$(5.6)\qquad \lim_{n\to\infty}\rho^{n+1}\left\|\sum_{t=0}^{\infty}\alpha^{t+1}P^t(R)\psi^{(n)}(\pi^{t+1})\right\|_{\mu} = 0$$

for $\rho < 1/c_2$.

The ψ's can be used to construct a partial Laurent series for $v^{\alpha,T}(R) := \Sigma_{t=0}^{T-1}\alpha^t P^t(R) r(\pi^{t+1})$ from which an expression for $v^{\alpha}(R)$ can be derived. Note that we use $\alpha = 1/(1+\rho)$ interchangeably.

THEOREM 5.3. *For any memoryless policy $R = (\pi^1, \pi^2, \ldots)$, all $N = 1,2,\ldots$ and $T = 1,2,\ldots$ it holds for ρ small enough that*

$$(5.7)\qquad v^{\alpha,T}(R)_l = \frac{1+\rho}{\rho}\left\{(1-\alpha^T)\hat{y}^{(-1)} + \sum_{t=0}^{T-2}(\alpha^{t+1}-\alpha^T)P^t(R)\psi^{(-1)}(\pi^{t+1})\right\}$$

$$+(1+\rho)\sum_{n=0}^{N}\rho^n\Big\{\hat{y}^{(n)} - \alpha^{T+n+1}P^{T+n}(R)\hat{y}^{(n)}$$

$$+\sum_{t=0}^{T+n-1}\alpha^{t+1}P^t(R)\psi^{(n)}(\pi^{t+1})\Big\} - \rho^{N+1}\sum_{t=0}^{T+N}\alpha^t P^t(R)\hat{y}^{(N)}$$

and that the limit for $N \to \infty$ can be taken yielding the relation,

$$(5.8)\qquad v^{\alpha}(R) = \frac{1+\rho}{\rho}\left\{\hat{y}^{(-1)} + \sum_{t=0}^{\infty}\alpha^{t+1}P^{t}(R)\psi^{(-1)}(\pi^{t+1})\right\}$$
$$+(1+\rho)\sum_{n=0}^{\infty}\rho^{n}\left\{\hat{y}^{(n)} + \sum_{t=0}^{\infty}\alpha^{t+1}P^{t}(R)\psi^{(n)}(\pi^{t+1})\right\}.$$

PROOF. The entire proof can be found in Dekker (1985, Chapter 1, Theorem 6.2). It is basically the same as that of Theorem 3.1 of Hordijk and Sladky (1977), the only difference being the fact that in the multichain case $\hat{y}^{(-1)}$ is no longer a constant vector. Therefore we replace $P^{t}(R)\hat{y}^{(-1)}$ by $\hat{y}^{(-1)} + \sum_{k=0}^{t-1}P^{k}(R)\psi^{(-1)}(\pi^{k+1})$, which follows from repeatedly applying (5.1). The rest of the proof remains the same as the necessary limits can be taken by (5.3), (5.4), (5.5), and (5.6). ■

We are now in a position to prove the main theorem of this section which is a generalization of §5 of Hordijk and Sladky (1977). Although no unichainedness assumption was used in that section, it did contain a rather unsatisfactory assumption. In the proof below we are able to skip this assumption by using the compactness of the class of decision rules and the continuity of $P(\pi)$ and $\psi^{(k)}(\pi)$ in π.

THEOREM 5.4. *Suppose Conditions* 5, 6 *hold. For any policy $R \in C_M$ we have*

$$(5.9)\qquad \limsup_{\rho\downarrow 0}\rho^{-n}\left[v_i^{\rho}(R) - \hat{y}_i\right] \leqslant 0, \qquad i \in E, \quad n = -1, 0, 1, \ldots$$

PROOF. Consider any $i \in E$ and $R = (\pi^1, \pi^2, \ldots), \in C_M$. By Theorem 5.3 we have

$$(5.10)\qquad v^{\rho}(R) - \hat{y}_{l} = (1+\rho)\sum_{n=-1}^{\infty}\rho^{n}\sum_{t=0}^{\infty}\alpha^{t+1}P^{t}(R)\psi^{(n)}(\pi^{t+1}).$$

Remark that for any decision rule π and $i \in E$ the sequence $\psi_i^{(-1)}(\pi), \psi_i^{(0)}(\pi), \ldots$ is lexicographic nonpositive.

Let $\phi^{(n)}(\alpha, R) = \sum_{t=0}^{\infty}\alpha^{t+1}P^{t}(R)\psi^{(n)}(\pi^{t+1})$. Note that by (5.4)

$$(5.11)\qquad \left\|\phi^{(n)}(\alpha, R)\right\|_{\mu} \leqslant \frac{c_1\tilde{c}}{1-\alpha}(c_2)^{n}, \qquad 0 < \alpha < 1, \quad R \in C_M.$$

Let $n_0 := \inf\{n \mid P_{ij}^{t}(R)\psi_j^{(n)}(\pi^{t+1}) \neq 0$ for some j and $t\}$.

If $n_0 = \infty$ then $v_i^{\rho}(R) - \hat{y}_{il} = 0$ which implies the assertion of the theorem. Otherwise there is some set JT such that

$$(5.12)\qquad P_{ij}^{t}(R)\psi_j^{(n)}(\pi^{t+1})\begin{cases} = 0, & t = 0, 1, \ldots,\ j \in E, \quad n < n_0, \\ < 0, & (j, t) \in JT, \quad n = n_0, \\ = 0, & (j, t) \notin JT, \quad n = n_0. \end{cases}$$

Hence we have

$$\rho^{-n_0}\left[v_i^{\rho}(R) - \hat{y}_i\right]_l = (1+\rho)\sum_{n=n_0}^{\infty}\rho^{n-n_0}\phi_i^{(n)}(\alpha, R).$$

By (5.4)

$$\lim_{\rho\downarrow 0}\rho^{n-n_0}\phi_i^{(n)}(\alpha, R) = 0 \quad \text{for } n \geqslant n_0 + 2 \quad \text{and}$$

$$\begin{aligned}\limsup_{\rho\downarrow 0}\rho^{-n_0}\left[v_i^{\rho}(R)-\hat{y}_i\right] &= \limsup_{\rho\downarrow 0}\left\{\phi_i^{(n_0)}(\alpha,R)+\rho\phi_i^{(n_0+1)}(\alpha,R)\right\}\\ &= \limsup_{\rho\downarrow 0}\sum_{t=0}^{\infty}\alpha^{t+1}\sum_j P_{ij}^t(R)\left\{\psi_j^{(n_0)}(\pi^{t+1})+\rho\psi_j^{(n_0+1)}(\pi^{t+1})\right\}\\ &\leqslant \sum_j \limsup_{\rho\downarrow 0}\sum_{t=0}^{\infty}\alpha^{t+1}P_{ij}^t(R)\left\{\psi_j^{(n_0)}(\pi^{t+1})+\rho\psi_j^{(n_0+1)}(\pi^{t+1})\right\}\end{aligned}$$

the latter step following from Fatou's lemma. We will show that for any $j\in E$ the lim sup is nonpositive.

Let $j\in E$ be fixed and note that $\rho\downarrow 0$, if and only if $\alpha\uparrow 1$. If

$$\limsup_{\alpha\uparrow 1}\sum_{t=0}^{\infty}\alpha^{t+1}P_{ij}^t(R)\psi_j^{(n_0)}(\pi^{t+1})=-\infty,$$

then the proof is obvious, since (cf. (5.4))

$$\limsup_{\alpha\uparrow 1}\rho\sum_{t=0}^{\infty}\alpha^{t+1}P_{ij}^t(R)\psi_j^{(n_0+1)}(\pi^{t+1})<\infty.$$

Otherwise we have by (5.12)

$$P_{ij}^t(R)\psi_j^{(n_0)}(\pi^{t+1})\to 0,\quad \text{as } t\to\infty. \tag{5.13}$$

In that case we consider $\limsup_{t\to\infty}P_{ij}^t(R)\psi_j^{(n_0+1)}(\pi^{t+1})$. Suppose it is positive, hence there exist a subsequence t_k, $k=1,2,\dots$ and an index k_0 such that for some $\varepsilon>0$

$$P_{ij}^{t_k}(R)\psi_j^{(n_0+1)}(\pi^{t_k+1})>\epsilon,\qquad k>k_0.$$

Since $\psi_j^{(n_0+1)}(\pi^{t_k+1})$ is bounded by some constant, say $\tilde{\tilde{c}}$, this implies that

$$P_{ij}^{t_k}(R)>\frac{\epsilon}{\tilde{\tilde{c}}}\quad\text{and}\quad \psi_j^{(n_0+1)}(\pi^{t_k+1})>\epsilon,\qquad k>k_0. \tag{5.14}$$

As by Assumption 1 the class of decision rules is a compact set, the sequence $\{\pi^{t_k+1}\}_{k=1}^{\infty}$ has a convergent subsequence, say $\pi^{t_{k'}+1}\to\pi_0$.

It then follows from (5.14) and the pointwise continuity of $\psi^{(n)}(\pi)$ for all n, that $\psi_j^{(n_0+1)}(\pi_0)\geqslant\epsilon$. Since by (5.12) and (5.14) $\psi_j^{(n)}(\pi^{t_{k'}+1})=0$ for all k' and $n<n_0$ and $\psi_j^{(n_0)}(\pi^{t_{k'}+1})\to 0$ by (5.13) we have also $\psi_j^{(n)}(\pi_0)=0$, $-1\leqslant n\leqslant n_0$. This implies that the first nonzero term of $\psi_j^{(-1)}(\pi_0),\dots,\psi^{(n_0)}(\pi_0)$, $\psi^{(n_0)}(\pi_0)$, $\psi_j^{(n_0+1)}(\pi_0),\dots$ is positive which is in contradiction with the fact that $\hat{y}$ is a solution to the Blackwell optimality equations. Hence we have

$$\limsup_{t\to\infty}P_{ij}^t(R)\psi_j^{(n_0+1)}(\pi^{t+1})\leqslant 0.$$

Denote $P_{ij}^t(R)\psi_j^{(n_0+1)}(\pi^{t+1})$ by $\xi_{ij}^t(R)$. We have

$$\limsup_{\alpha\uparrow 1}(1-\alpha)\sum_{t=0}^{\infty}\alpha^t\xi_{ij}^t(R)\leqslant\limsup_{\alpha\uparrow 1}(1-\alpha)\sum_{t=0}^{\infty}\alpha^t\Big(\sup_{k\geqslant t}\xi_{ij}^k(R)\Big).$$

Since $\lim_{t\to\infty}(\sup_{k\geqslant t}\xi_{ij}^k(R))$ exists, it is equal to the right-hand side of the above expression, which follows from the regularity of the Abelian summation method (i.e., every convergent series is summed to its normal sum). Hence we have

$$\limsup_{\alpha\uparrow 1}(1-\alpha)\sum_{t=0}^{\infty}\alpha^{t}\Big(\sup_{k\geqslant t}\xi_{ij}^{k}(R)\Big)=\limsup_{t\to\infty}\xi_{ij}^{t}(R)\leqslant 0$$

and accordingly,

$$\limsup_{\alpha\uparrow 1}\frac{(1-\alpha)}{\alpha}\sum_{t=0}^{\infty}\alpha^{t+1}P_{ij}^{t}(R)\psi_{j}^{(n_0+1)}(\pi^{t+1})\leqslant 0.$$

Replacing $(1-\alpha)/\alpha$ by ρ now yields

$$\limsup_{\alpha\uparrow 1}\sum_{t=0}^{\infty}\alpha^{t+1}P_{ij}^{t}(R)\left\{\psi_{j}^{(n_0)}(\pi^{t+1})+\rho\psi_{j}^{(n_0+1)}(\pi^{t+1})\right\}\leqslant 0,$$

which completes the proof. ■

Let $v\in LS_{\mu}^{E}$ and define $\tilde{\psi}^{(-1)}(\pi),\tilde{\psi}^{(0)}(\pi),\ldots$ similar to (5.1) with $\hat{y}$ replaced by v. Suppose that for a policy R the sequence $\tilde{\psi}_{i}^{(-1)}(\pi),\tilde{\psi}_{i}^{(0)}(\pi),\ldots$ is lexicographically nonpositive for each $i\in E$, for each decision rule π prescribed by R and *each* cluster point of its decision rules, then the entire proof of Theorem 5.4 can be used to show that

$$\limsup_{\rho\downarrow 0}\rho^{-n}\big(v_{i}^{\rho}(R)-v\big)\leqslant 0,\qquad n=-1,0,\ldots,\ i\in E.$$

If v^{ρ} is contained in LS^{E}, then we would have $v^{\rho}(R)\leqslant_{l} v$ and we could establish an extension of Condition 3 to nonstationary policies. However, in general, $v^{\rho}(R)$ does not have a Laurent series expansion.

A second remark is on n-discount optimality. The proof of Theorem 5.4 can be used to show that the n-discount optimality equations are constituted by $U^{(-1)}y=0,\ldots,U^{(n+1)}y=0$. For, suppose that $\hat{y}$ is a solution to these equations and define vectors $\psi^{(-1)(\pi)},\ldots,\psi^{(n+1)}(\pi)$ as in (5.1). Hence for every decision rule π and state $i\in E$ the sequence $\psi_{i}^{(-1)}(\pi),\ldots,\psi_{i}^{(n+1)}(\pi)$ is lexicographic nonpositive. From the proof of Theorem 5.4 it follows that for any policy R, $\limsup_{\rho\downarrow 0}\rho^{-n}[v_{i}^{\rho}(R)-\hat{y}_{i}]\leqslant 0$, $i\in E$ and that for a persistently conserving policy (i.e., $\psi^{(k)}(\pi^{t})=0$ for $k=-1,\ldots,n+1$ and all t) the equality sign holds.

6. Examples. In this section we present two types of models for which our analysis is suited. Both types have a multichain structure. The μ-uniform geometric convergence condition is in both cases verified for some vector μ; verification of Conditions 5 and 6 is essentially not more difficult and releases an aperiodicity assumption.

Remark that in an aperiodic, finite-state and finite-action Markov decision chain the uniform geometric convergence condition holds for some constants $c>0$ and $\beta<1$. In that case $P^{k}(f)\to\Pi(f)$, for $k\to\infty$, pointwise on E and finiteness of F implies that $\|P^{n_0}(f)-\Pi(f)\|_{e}\leqslant 1-\epsilon$, for some integer $n_0>0$ and $\epsilon>0$ from which the uniform geometric convergence follows.

The first type of models is an extension of the aperiodic finite-state and finite action decision chain.

Type 1. The state space E can be written as $E=E_0\cup E_1\cup\cdots\cup E_N$ (N may be ∞), whereby E_k, $k=1,2,\ldots,N$ are finite closed and aperiodic sets under all policies and on each E_k the sets of available actions are also finite.

Only if $N = \infty$, we have to assume that there exist constants $c_1 > 0$ and $\beta_1 < 1$ such that condition e-UGC(c_1, β_1) holds on E_k, for $k = 1, 2, \ldots$. (If $N < \infty$, existence of c_1 and β_1 requires no assumption.) The set E_0 may be denumerable. On E_0 we assume a vector $\tilde{\mu} > 0$ such that for constants $c_0, \tilde{c} > 0$, $\tilde{\beta}_0 < 1$ and $n_0 \in \mathbb{N}$

$$(6.1) \qquad \|P(f)\|_{\mu} \leqslant \tilde{c}, \qquad f \in F,$$

$$(6.2) \qquad \|_{E_0^c}P^{n_0}(f)\|_{\tilde{\mu}} \leqslant \tilde{\beta}_0, \qquad f \in F,$$

$$(6.3) \qquad \|r(f)\|_{\mu} \leqslant c_0, \qquad f \in F,$$

where μ equals $\tilde{\mu}$ on E_0 and equals e on E_0^c and

$$_{E_0^c}P_{ij}(f) = \begin{cases} 0 & \text{for } j \in E_0^c \\ P_{ij}(f) & \text{for } j \in E_0 \end{cases}, \quad i \in E,\ f \in F.$$

The n-fold matrix product of $_{E_0^c}P(f)$ with itself is denoted as $_{E_0^c}P^n(f)$. Remark that by (6.1) and (6.2) we have

$$(6.4) \qquad \|_{E_0^c}P^m(f)\|_{\mu} \leqslant \bar{c}_1\beta_0^m, \qquad m = 1, 2, \ldots, f \in F,$$

for some constants $\bar{c}_1 > 0$ and $\beta_0 < 1$.

To verify μ-uniform convergence we have to consider

$$\frac{1}{\mu_i}\sum_j |P_{ij}^n(f) - \Pi_{ij}(f)|\mu_j.$$

From the foregoing it is clear that for $i \in E_0^c$ this expression is smaller than $c_1(\beta_1)^n$, uniformly for all $f \in F$. For $j \in E_0$ consider $\Pi_{ij}(f)$. Suppose it is positive, which implies that

$$\lim_{n\to\infty} \frac{1}{n+1}\sum_{k=0}^{n} P_{ij}^k(f) > 0,$$

hence $\limsup_{n\to\infty} P_{ij}^n(f) > 0$. Remark that $P_{ij}^n(f) = {}_{E_0^c}P_{ij}^n(f)$ for $j \in E_0$, since E_0^c is closed.

Hence

$$\limsup_{n\to\infty} \frac{1}{\mu_i}\, {}_{E_0^c}P_{ij}^n(f)\mu_j > 0,$$

which contradicts (6.4) and accordingly $\Pi_{ij}(f) = 0$, $i, j \in E_0$, $f \in F$.

With respect to $P_{ij}^n(f)$ we apply a first entrance decomposition into E_0^c. For $j \in E_0^c$ it can easily be proven that

$$P_{ij}^n(f) = \sum_{l=0}^{n-1}\sum_{m\in E_0}\sum_{h\in E_0^c} P_{im}^l(f)P_{mh}(f)P_{hj}^{n-l-1}(f).$$

Since $\Pi(f) = P^n(f)\Pi(f)$ we can prove in a similar way that

$$\Pi_{ij}(f) = \sum_{l=0}^{n-1}\sum_{m\in E_0}\sum_{h\in E_0^c} P_{im}^l(f)P_{mh}(f)\Pi_{hj}(f).$$

Hence by (6.4) and (6.1) we have for $i \in E_0$,

$$\begin{aligned}\frac{1}{\mu_i}\sum_j \left|P_{ij}^n(f) - \Pi_{ij}(f)\right|\mu_j &= \frac{1}{\mu_i}\sum_{j\in E_0} P_{ij}^n(f)\mu_j \\ &\quad + \sum_{l=0}^{n-1}\sum_{m\in E_0}\frac{1}{\mu_i}P_{im}^l(f)\mu_m \sum_{h\in E_0^c}\frac{P_{mh}(f)\mu_h}{\mu_m} \\ &\quad \times \sum_{j\in E_0^c}\frac{1}{\mu_h}\left|P_{hj}^{n-l-1}(f) - \Pi_{hj}(f)\right|\mu_j \\ &\leqslant \bar{c}_1\beta_0^n + \sum_{l=0}^{n-1}\bar{c}_1\beta_0^l\tilde{c}c_1\beta_1^{n-l-1} \leqslant n\bar{c}\bar{\beta}^n\end{aligned}$$

for some constant $\bar{c}$, $\bar{\beta} = \max(\beta_0, \beta_1)$ and for all $i \in E_0$, $f \in F$ and $n = 1, 2, \ldots$. This implies that for some other constants c, β the expression is smaller than $c\beta^n$. From the above it is clear the Markov decision chains of this type satisfy the μ-uniform convergence condition with bounding vector μ and constants c and β, which, together with (6.3), is sufficient to establish Blackwell optimality equations.

Remark that if $r(f)$ is bounded on E and $\mu = e$, condition (6.2) reduces to

$$\sum_{j\in E} {}_{E_0^c}P_{ij}^{n_0}(f) \leqslant 1 - \epsilon, \qquad f \in F, i \in E,$$

for some $\epsilon > 0$, which is the simultaneous Doeblin condition in case $\Sigma_j P_{ij}(f) = 1$.

The second type of models is a class of optimal stopping problems. It shows that our conditions are also applicable for total reward criteria.

Type 2. Let $E = \mathbb{Z}$.

Transition probabilities:

action 0 or stopping action: $P_{ii}(0) = 1$, $i \in E$,

action 1 or nonstopping action: $0 \leqslant P_{ii+1}(1) \leqslant p$, $0 < q \leqslant P_{ii-1}(1) \leqslant 1$, $P_{ii}(1) = 1 - P_{ii+1}(1) - P_{ii-1}(1)$, $i \in E$, for fixed $p \leqslant q$.

Let $\mu_i = (1 + y)^i$, $i \in E$ for some $y > 0$.

Remark that $\|P(0)\|_\mu = 1$ and that $\|P(1)\|_\mu \leqslant 1 + py - qy/(1 + y)$ ($P(0)$ and $P(1)$ are the transition probability matrices when in all states action 0, 1 respectively are chosen). Hence for fixed $p < q$ there exist some y_0 and a β_0 such that $\|P(1)\|_\mu \leqslant \beta_0 < 1$. Assume with respect to the immediate rewards that

$$\sup_{i\in E}\ \sup_{a=0,1}\frac{|r_i(a)|}{(1 + y_0)^i} < \infty.$$

Consider any $f \in F$ and let $E_0(f) = \{i \in E \mid f(i) = 1\}$, implying that $f(i) = 0$ on $E_0^c(f)$ and that every state of $E_0^c(f)$ is absorbing.

Note that regardless of f we have $\|_{E_0^c(f)}P(f)\|_\mu \leqslant \beta_0$ and $\|P(f)\|_\mu = 1$. Hence conditions (6.1) and (6.2) of the previous example are fulfilled with constants independent of f! Remark that on $E_0^c(f)$ we have $P(f) = \Pi(f)$ and therefore the constants c_1, β_1 of the previous example are zero. After inserting this into the results of the previous example, we arrive at

$$\|P^n(f) - \Pi(f)\|_\mu \leqslant \beta_0^n \quad \text{for all } f \in F.$$

This implies that the Markov decision chain is μ-uniform geometric convergent with constants $c_0 = 1$ and β_0, which is sufficient for Blackwell optimality. It will be obvious that some of the assumptions of this example can be relaxed.

Another example of a multichain structure can be found in Dekker (1985), where the state space is multi-dimensional.

References

Bather, J. A. (1973). Optimal Decision Procedures for Finite Markov Chains. Part i. Examples. *Adv. in Appl. Probab.* **5** 328–339. Part ii. Communicating Systems. *Adv. in Appl. Probab.* **5** 521–540. Part iii. General Convex Systems. *Adv. in Appl. Probab.* **5** 541–553.

Blackwell, D. (1962). Discrete Dynamic Programming. *Ann. Math. Statist.* **33** 719–726.

______. (1965). Discounted Dynamic Programming. *Ann. Math. Statist.* **36** 226–235.

______. (1967). Positive Dynamic Programming. in: *Proc. 5th Berkeley Sympos. Math. Statistics and Probability.* Vol. I. 415–418.

Chung, K. L. (1960). *Markov Chains with Stationary Transition Probabilities.* Springer Verlag, Berlin.

Dekker, R. (1985). Denumerable Markov Decision Chains: Optimal Policies for Small Interest Rates. Ph.D. thesis, Univ. of Leiden.

Denardo, E. V. (1967). Contraction mappings in the Theory Underlying Dynamic Programming. *Siam Rev.* **9** 165–177.

______ and Miller, B. L. (1968). An Optimality Condition for Discrete Dynamic Programming with no Discounting. *Ann. Math. Statist.* **39** 1220–1227.

Deppe, H. (1984). On the Existence of Average Optimal Policies in Semiregenerative Decision Models. *Math. Oper. Res.* **9** 558–575.

Derman, C. (1966). Denumerable State Markovian Decision Processes—Average Cost Criterion. *Ann. Math. Statist.* **42** 1545–1553.

______. (1970). *Finite State Markovian Decision Processes.* Academic Press, New York.

______ and Veinott, A. F., Jr. (1967). A Solution to a Countable System of Equations Arising in Markovian Decision Processes. *Ann. Math. Statist.* **38** 582–584.

Dietz, H. M. and Nollau, V. (1983). *Markov Decision Problems with Countable State Spaces.* Band 15, Akademie Verlag, Berlin.

Federgruen, A., Hordijk, A. and Tijms, H. C. (1978). A Note on the Simultaneous Recurrence Conditions on a Set of Denumerable Stochastic Matrices. *J. Appl. Probab.* **15** 842–847.

______, ______ and ______. (1979a). Denumerable State Semi-Markov Decision Processes with Unbounded Costs, Average Cost Criterion. *Stochastic Process Appl.* **9**, 223–235.

______, ______ and ______. (1979b). Recurrence Conditions in Denumerable State Markov Decision Processes. In: M. L. Puterman (Ed.), *Dynamic Programming and its Applications.* Academic Press, New York.

______, Schweitzer, P. J. and ______. (1983). Denumerable Undiscounted Semi-Markov Decision Processes with Unbounded Rewards. *Math. Oper. Res.* **8** 298–313.

Hordijk, A. (1972). On Doeblin's Condition and Its Application in Markov Decision Processes. Math. Centre Report BW 15/72, Amsterdam (in Dutch).

______. (1974). Dynamic Programming and Markov Potential Theory. Math. Centre Tract no. **51**, Amsterdam.

______. (1976). Regenerative Markov Decision Models. In: R. J. B. Wets (Ed.), *Math. Programming Study* **6**, North-Holland, Amsterdam.

______ and Dekker, R. (1983). Denumerable Markov Decision Chains: Sensitive Optimality Criteria. *Operations Research Proceedings* 1982. Springer-Verlag, Berlin.

______ and Sladky, K. (1977). Sensitive Optimality Criteria in Countable State Dynamic Programming. *Math. Oper. Res.* **2** 1–14.

Kadota, Y. Countable State Markovian Decision Processes under the Doeblin-Condition. *Res. Assoc. Statist. Sci.* **19** 85–94.

Mann, E. (1985). Optimality Equations and Sensitive Optimality in Bounded Markov Decision Processes. *Optimization* **16** 767–781.

______. (1986). Optimalitätsgleichungen für undiskontierte semi-Markoffsche Entscheidungsprozesse. Ph.D.-thesis, Univ. of Bonn.

Miller, B. L. and Veinott, A. F., Jr. (1969). Discrete Dynamic Programming with a Small Interest Rate. *Ann. Math. Statist.* **40** 366–370.

Ross, S. M. (1968). Non-Discounted Denumerable Markovian Decision Models. *Ann. Math. Statist.* **39** 412–423.

Schäl, M. (1987). Paper presented at the Oberwolfach meeting on Operations Research.

Schweitzer, P. J. (1982). Solving MDP Functional Equations by Lexicographic Optimization. RAIRO **16** 91–98.

Sennott, L. I. (1986). A New Condition for the Existence of Optimal Stationary Policies in Average Cost Markov Decision Processes. *Oper. Res. Letters* **5** 17–23.

Syski, R. (1978) Ergodic Potential. *Stochastic Process Appl.* **7** 311–336.

Tijms, H. C. (1975). On Dynamic Programming with Arbitrary State Space, Compact Action Space and the Average Return as Criterion. Research Report BW 55/75, Math. Centre, Amsterdam.

Titchmarsh, E. C. (1939). *The Theory of Functions*. Oxford University Press, London.

Veinott, A. F., Jr, (1969). On Discrete Dynamic Programming with Sensitive Discount Optimality Criteria. *Ann. Math. Statist.* **40**, 1635–1660.

______. (1974). Markov Decision Chains. in: G. B. Dantzig and B. C. Eaves (Eds.), *Studies in Math. Vol.* 10: *Studies in Optimization*. Math. Assoc. of America, 124–159.

Wijngaard, J. (1977). Stationary Markovian Decision Problems and Pertubation Theory of Quasi-Compact Linear Operators. *Math. Oper. Res.* **2** 91–102.

Zijm, W. H. M. (1984). The Optimality Equations in Multichain Denumerable State Markov Decision Processes with the Average Cost Criterion: The Unbounded Cost Case. C. Q. M. Note 22, Centre for Quant. Methods, Philips B. V. Eindhoven.

______. (1985). The Optimality Equations in Multichain Denumerable State Markov Decision Processes with the Average Cost Criterion: The Bounded Cost Case. *Stat. & Decisions*.

DEKKER: KONINKLIJKE/SHELL LABORATORIUM AMSTERDAM, P.O. BOX 3003, AMSTERDAM, THE NETHERLANDS

HORDIJK: INSTITUTE OF APPLIED MATHEMATICS AND COMPUTER SCIENCE, UNIVERSITY OF LEIDEN, LEIDEN, THE NETHERLANDS