MATHEMATICS OF OPERATIONS RESEARCH
Vol. 8, No. 2, May 1983
Printed in U.S.A.

DENUMERABLE UNDISCOUNTED SEMI-MARKOV DECISION PROCESSES WITH UNBOUNDED REWARDS*

A. FEDERGRUEN,† P. J. SCHWEITZER‡ AND H. C. TIJMS§

This paper establishes the existence of a solution to the optimality equations in undiscounted semi-Markov decision models with countable state space, under conditions generalizing the hitherto obtained results. In particular, we merely require the existence of a *finite* set of states in which every pair of states can reach each other via some stationary policy, *instead of* the traditional and restrictive assumption that every stationary policy has a *single irreducible* set of states. A replacement model and an inventory model illustrate why this extension is essential. Our approach differs fundamentally from classical approaches; we convert the optimality equations into a form suitable for the application of a fixed point theorem.

1. Introduction. This paper establishes conditions for the existence of a solution to the countable set of functional equations

$$v_i^* = \sup_{a \in A(i)} \left\{ r(i,a) - g^* \tau(i,a) + \sum_{j \in I} p_{ij}(a) v_j^* \right\}, \qquad i \in I. \tag{1}$$

These equations arise in undiscounted Markov Renewal Programs (MRPs) with a *denumerable* state space. Here, at epochs beginning with epoch 0, a system is observed to be in one of the states of a denumerable state space I and is subsequently controlled by choosing an action. For any state $i \in I$, the set $A(i)$ denotes the set of pure actions available in state i. When choosing action $a \in A(i)$ in state i, the following two consequences are incurred, regardless of the decision epoch at which this action is taken or the previous history of the system:

(i) An immediate expected reward $r(i,a)$ is incurred.

(ii) The state of the system at the next decision epoch is $j \in I$ with probability $p_{ij}(a)$; the expected holding time in state i is denoted by $\tau(i,a) < \infty$. Observe that $\sum_{j \in I} p_{ij}(a) = 1$ for all i, a.

Establishing the existence of a solution to the functional equations (1) is essential for proving the existence of a stationary policy which is optimal under the *strong* average return per unit time criterion (cf. §2):

(i) with *bounded* one-step rewards, i.e., in case $|r(i,a)| \leqslant M$ for some $M > 0$ and all $i \in I$, $a \in A(i)$, Ross [22, Theorem 7.6] exhibits that any policy which chooses for all $i \in I$ an action achieving the supremum at the right hand side of (1), is average return optimal (in the strong sense);

(ii) with *unbounded* one-step expected rewards Federgruen et al. [11] showed that the existence of an average return optimal policy follows from the existence of a solution to the optimality equation (1) provided this solution satisfies some additional regularity condition (cf. Assumption 5).

*Received November 12, 1980; revised March 19, 1982.
AMS 1980 subject classification. Primary: 90C45.
OR/MS Index 1978 subject classification. Primary: 118 Semi-Markov dynamic programming.
Key words. Semi-Markov decision models, countable state space, average return per unit time, fixed point approach, recurrency conditions.
†Columbia University.
‡University of Rochester.
§Vrije Universiteit, Amsterdam.

0364-765X/83/0802/0298$01.25

It is well known that an average return optimal policy does not need to exist; several examples have been given of the occurrence of possible irregularities (cf., e.g., Ross [22], Fisher and Ross [15]). Federgruen et al. [11], generalizing the hitherto obtained results, established a rather complete theory for this denumerable state semi-Markov decision model with unbounded one-step expected rewards under the *unichainedness assumption* where the *transition probability matrices* (tpm's) of *the stationary policies* each have a *single* ergodic set. In fact this assumption of all transition probability matrices being unichained is common to the entire literature on the denumerable state space model with the exception of some of the results in Hordijk [17] (cf., e.g., Theorem 11.8) and Wijngaard [33].

In this paper we extend the existence conditions for a solution to the optimality equation (1) to systems satisfying a recurrency condition which is substantially weaker than the above mentioned unichainedness assumption: we require the existence of a finite set of states in which every pair of states can reach each other via some policy (the assumption extends the communicatingness condition in *finite* MRPs, as introduced by Bather [1]). Similarly, our results include Hordijk's [17] communicating systems with *bounded* parameters, as special cases. The additional conditions needed are similar to the additional conditions employed in Federgruen et al. [11]. Roughly speaking we require (i) the one step expected rewards to be bounded from above though not necessarily from below; (ii) the existence of a finite set of states such that the supremum over all stationary policies of the expected time, as well as the total absolute rewards until the first visit to this set is finite for every starting state; (iii) the expectation of the value of these suprema in the state adopted at the first transition is continuous in the action taken in the starting state. ((iii) is trivially met in models with *finite* action spaces.)

Similar conditions were recently obtained in Deppe [3] to prove the existence of policies which are optimal in the *weak* sense (cf. §2).

The new conditions thus encompass a number of hitherto intractable models.

EXAMPLE 1. We consider the Blast Furnaces problem, discussed by Stengos and Thomas [31] as well as Stengos [30, Chapter 5]. One has n identical pieces of equipment which, from time to time, will need overhauling, either because they have failed during operation or to prevent such a failure. A piece of equipment which has completed its last overhaul i periods ago will survive the next period with probability p_i. An overhaul takes L_1 periods of time. When m pieces of equipment ($1 \leqslant m \leqslant n$) are overhauled, the loss of revenue to the system is given by a convex function $c(m)$. In [31] no operating costs are considered; although multichain policies may occur (as explained below) the absence of operating costs allows for a restriction of the action sets which excludes such policies. In fact the restricted model in [31] has a bounded cost structure, and can easily be shown to satisfy the simultaneous Doeblin condition which implies the existence of a bounded solution to the optimality equation (cf. Hordijk [17] and Federgruen and Tijms [8]). In the presence of an operating cost function $h(i)$ (for each piece of equipment) where i denotes the age of the piece since its last overhaul, the above restriction of the policy space cannot be applied. In fact multichain policies may even be optimal and to our knowledge only the conditions in this paper can be used to verify the existence of a solution to the optimality equations as well as the existence of a stationary optimal policy. Appendix 1 verifies Assumptions 1–5 for the special case $n = 2$.

EXAMPLE 2. Silver [28], Silver and Thompstone [29] and Federgruen et al. [9] consider a continuous review multi-item inventory system where the demand processes for the items are independent compound Poisson processes (the demand size is a nonnegative, integer random variable). There is a major setup cost associated with a replenishment of the family. For each individual item included in the replenishment,

an item specific (minor) setup cost is added. In addition, the cost structure consists of holding-, shortage- and variable replenishment costs. Excess demands are backlogged and every order has a fixed lead time. The solution methods in [9], [28] and [29] decompose the coordinated control problem into single-item problems for each item in the family. Each single-item problem has "normal" replenishment opportunities (at the *major* setup cost) occurring at the demand epochs for this item and "*special*" replenishment opportunities (at *reduced* setup costs) at epochs generated by a Poisson process approximating the superposition of the ordering processes triggered by other items.

At replenishment opportunities, the system may not be left with more than L units short (say), or with a positive net inventory position = (inventory on hand) + (outstanding orders) − (backlog) of more than U units (say).

For these single item models, conditions 1–5 for the existence of a stationary optimal policy are easily verified with a convenient choice of the set K (see Appendix 2).

In addition, the approach taken to prove the existence of a solution to the optimality equation (1) is altogether different from the approaches in the models where *unichainedness* is assumed. There the existence of a solution to (1) is obtained from the limiting behavior of the total maximal discounted return vector, as the discount factor tends to one (cf. [4], [8], [10], [17], [18], [22] and [32]), using a technique introduced by Taylor [32] and Ross [22]. Here, we convert the equation into a form suitable for the application of the Tychonoff fixed point theorem which is a generalization of the well-known Brouwer fixed point theorem (cf. [7]). The same approach has been used in Federgruen and Schweitzer [13] to establish a simple existence proof for a solution to the optimality (vector) equations that arise in the *general* model with *finite* state and action spaces, where the maximal gain rate vector may have unequal components. Finally, our approach enables a (partial) characterization of the optimality equation's solution set.

We conclude this introduction by pointing out the plan of the paper. In §2 we give some notation and preliminary results. In §3 the existence of a solution to optimality equation (1) is obtained; also, the solution set of (1) is characterized.

2. Preliminaries and notation. We first make the following assumption with respect to the parameters and action sets in our model:

ASSUMPTION 1 (*cf.* [11]). (a) *For any $i \in I$, the set $A(i)$ is a compact metric set*; (b) *for any $i \in I$, $r(i,a)$ is upper-semicontinuous on $A(i)$ and $\tau(i,a)$ and, for all $j \in I$, $p_{ij}(a)$ are continuous on $A(i)$*; (c) *there is a number $\epsilon > 0$ such that $\tau(i,a) \geqslant \epsilon$ for all $i \in I$ and $a \in A(i)$.*

With respect to the one-step expected rewards, we assume upper-semicontinuity rather than the more conventional *continuity* assumption to include, e.g., "cost or reward structures" with fixed components.

We next introduce some familiar notions. For $n = 0, 1, \ldots$ denote by X_n and a_n the state and the action at the nth decision epoch (the 0th decision epoch is at time 0). A policy π for controlling the system is any measurable rule which for each n specifies which action to choose at the nth decision epoch given the current state X_n and the sequence $(X_0, a_0, \ldots, X_{n-1}, a_{n-1})$ of past states and actions where the actions chosen may be randomized. A policy π is called *memoryless* when the actions chosen are independent of the history of the system except for the present state. Define $\mathcal{P}$ as the class of all stochastic matrices $P = (p_{ij})$, $i, j \in I$ such that for any $i \in I$ the elements of the ith row of P can be represented by

$$p_{ij} = \int_{A(i)} p_{ij}(a)\pi_i(da) \qquad \text{for all} \quad j \in I$$

for some probability distribution $\pi_i(\cdot)$ on $A(i)$. Then any memoryless policy π can be represented by some sequence $(P_1, P_2, \dots)$ in $\mathcal{P}$ such that the ith row of P_n gives the probability distribution of the state at the nth decision epoch when the current state at the $(n-1)$st decision epoch is i and policy π is used. Define $F = X_{i\in I} A(i)$; for any $f \in F$, let $P(f)$ be the stochastic matrix whose (i,j)th element is $p_{ij}(f(i))$, $i, j \in I$ and for $n = 1, 2, \dots$ denote by the stochastic matrix $P^n(f) = (p_{ij}^n(f))$ the n-fold matrix product of $P(f)$ with itself. For $n \geqslant 1$, let $\pi_{ij}^{(n)}(f) = \{p_{ij}^1(f) + \cdots + p_{ij}^n(f)\}/n$. It is well known from Markov chain theory (cf. Chung [2]) that the sequence $\{\pi_{ij}^{(n)}(f)\}_{n=1}^{\infty}$ has a limit $\pi_{ij}(f)$ (say), for all $i, j \in I$. A memoryless policy $R = (P_1, P_2, \dots)$ is called *randomized stationary* when $P_n = \psi$ for all $n \geqslant 1$ for some $\psi \in \mathcal{P}$, and is denoted by $\psi^{(\infty)}$. In the special case where $P_n = P(f)$ for all $n \geqslant 1$ and some $f \in F$, the policy is called (*pure*) *stationary* and denoted by $f^{(\infty)}$, since prescribing the single action $f(i) \in A(i)$ whenever in state i. Observe that under any (randomized) stationary policy the process $\{X_n\}_{n=1}^{\infty}$ is a Markov chain. A policy R^* is called average optimal in the strong sense when

$$\liminf_{n\to\infty} \frac{E_{R^*}\{\sum_{k=0}^n r(X_k, a_k) \mid X_0 = i\}}{E_{R^*}\{\sum_{k=0}^n \tau(X_k, a_k) \mid X_0 = i\}} \geqslant \limsup_{n\to\infty} \frac{E_R\{\sum_{k=0}^n r(X_k, a_k) \mid X_0 = i\}}{E_R\{\sum_{k=0}^n \tau(X_k, a_k) \mid X_0 = i\}} \tag{2}$$

for all $i \in I$ and policies R, where E_R denotes the expectation under policy R.

Especially when a solution to the optimality equation cannot be established, one often considers the *weak* variant of criterion (2) where the lim inf or the lim sup is used on both sides of the inequality in (2). However, the relations (2)–(4) in Flynn [16] show that these criteria are essentially weaker than the criterion (2).

In either sense, it is well known that an average return optimal policy does not need to exist even under very strong regularity conditions (cf. [4], [8], [10]–[12], [17]–[19], [24], [32] and [33]). In general we can only state that for a fixed initial state we may restrict ourselves to the memoryless policies (cf. Derman and Strauch [5]).

We now introduce our main assumption. First, for any stochastic process $\{X_n\}_{n=1}^{\infty}$ and $A \subseteq I$, define $N(A) = \inf\{n \geqslant 1 \mid X_n \in A\}$ where $N(A) = \infty$ if $X_n \notin A$ for all $n \geqslant 1$, i.e., $N(A)$ denotes the number of transitions until the first visit to the set A. Also for any $A \subseteq I$, and $\psi \in \mathcal{P}$, define for $i \in I$, and $n \geqslant 1$ the taboo probability

$$_A p_{ij}^n(\psi) = \text{Prob}_{\psi(\infty)}\{X_n = j, X_k \notin A \text{ for } 1 \leqslant k \leqslant n-1 \mid X_0 = i\}. \tag{3}$$

Observe that the expected first passage time from state i to set A under policy $\psi^{(\infty)}$, $\psi \in \mathcal{P}$ is given by:

$$\mu_{iA}(\psi) \overset{\text{def}}{=} E_{\psi^{(\infty)}}\{N(A) \mid X_0 = i\} = 1 + \sum_{n=1}^{\infty} \sum_{j\notin A} {}_A p_{ij}^n(\psi). \tag{4}$$

(For $A = \{j\}, j \in I$, we use the shorthand notation $\mu_{ij}(\psi)$.) A state $i \in I$ is said to *reach* state $j \in I$ if for some $f \in F$, $p_{ij}^n(f) > 0$ for some $n \geqslant 0$. A state $i \in I$ is called *positive recurrent* under $P(f)$ if $\mu_{ii}(f) < \infty$.

ASSUMPTION 2. (a) *There is a finite set K such that for any $i \in I$ the quantities $u^*(i)$ and $y^*(i)$, $i \in I$, are finite where*

$$u^*(i) = \sup_{f\in F} E_{f(\infty)}\left\{\sum_{k=0}^{N(K)-1} \tau(X_k, a_k) \middle| X_0 = i\right\} < \infty, \qquad i \in I, \tag{5}$$

$$y^*(i) = \sup_{f\in F} E_{f(\infty)}\left\{\sum_{k=0}^{N(K)-1} \middle| r(X_k, a_k)\middle| \,\middle| X_0 = i\right\} < \infty, \qquad i \in I. \tag{6}$$

(b) *Let* $K^* = \{i \in K \mid i$ *is positive recurrent under some* $P(f), f \in F\}$. *Then all states in* K^* *can reach each other.*

(Note that $u^*(i)$ and $y^*(i)$ are *not* assumed to be bounded in i, e.g., $u^*(i) = i$ is acceptable. This permits us to handle unbounded rewards. The subsequent analysis shows that when identifying a set K satisfying assumptions 2(a) and 3(a) below, *only* states that are *positive recurrent* under some stationary policy need to be included in this set; in other words, we can always choose $K = K^*$. In practical applications, however, it may be easier to verify (5) and (6) when including in the set K some states that are nonrecurrent under every stationary policy.)

As pointed out in the introduction, Assumption 2(b) constitutes the main extension of our existence conditions as compared to the unichainedness assumption in [11] as well as the vast majority of the literature. Our condition is related to the "communicatingness" condition in Bather [1] for models with *finite* state spaces. This condition calls for the existence of a (randomized) policy under which all states are recurrent (cf. also [27]). In fact, under Assumptions 1–2, the policies in F generate a collection of Markov chains, embedded on visits to K^*, which constitutes a communicating system in the sense of Bather [1], cf. Lemma 3 part (b) below. Assumptions 1–2 also imply that the MRP is K-communicative in Wijngaard's sense [33] (the latter is one of the existence conditions in [33]; we are unaware of examples where this property, which applies to Markov renewal processes embedded on visits to K, is directly verifiable). Hordijk [17] uses communicatingness conditions which are stronger than the one in [33]; cf. Deppe [3, p. 111] for a discussion of the relationships between the latter.

By Assumption 2(a) and the fact that $\tau(i,a) \geqslant \epsilon$ for all i,a, we have

$$m^*(i) = \sup_{f \in F} \mu_{iK}(f) < \infty \qquad \text{for all} \quad i \in I. \tag{7}$$

LEMMA 1. *Let Assumptions* 1–2 *hold*: (a) *Fix* $\psi \in \mathcal{P}$. *Let* C *be an irreducible set of states under* $\psi^{(\infty)}$. *The states of* C *are positive recurrent under* $\psi^{(\infty)}$ *and* $C \cap K = C \cap K^* \neq \emptyset$. *In particular, the set* K^* *is not empty.*

(b) *There exists a randomized stationary policy* $\psi^{*(\infty)}, \psi^* \in \mathcal{P}$ *under which the Markov chain* $\{X_n\}_{n=1}^{\infty}$ *has a single irreducible set of states. The set contains* K^* *and consists of positive recurrent states.*

PROOF. (a) In view of [11, (7) and Theorem 3] we have $m^*(i) = \sup_{a \in A(i)}\{1 + \sum_{j \notin K} p_{ij}(a) m^*(j)\}$ for all $i \in I$. In particular, $1 + \sum_{j \notin K} P_{ij}(\psi(i)) m^*(j) \leqslant m^*(i)$. This and the finiteness of K show that every irreducible set of states under $\psi^{(\infty)}$ is positive recurrent (cf. Hordijk [17, p. 53] and Chung [2, Theorem 3, p. 47]). (b) In view of assumption 2(b) there exists for each ordered pair (i, j) with $i, j \in K^*$, a stationary policy $f_{(i,j)}^{(\infty)}$ under which i reaches j. Take ψ^* as a finite mixture of $\{f_{(i,j)}^{(\infty)} \mid i, j \in K^*\}$. Let C be a closed irreducible set of states under $\psi^{*(\infty)}$. In view of Part (a), $C \cap K^* \neq \emptyset$. This, together with the fact that all states in K^* reach each other under $\psi^{*(\infty)}$, proves part (b) of the lemma. ∎

We now fix a policy $\psi^{*(\infty)}$ satisfying the properties in Lemma 1. This policy will play a crucial role in the verification of the existence of a solution to the optimality equation. To complete our list of assumptions, an additional requirement has to be imposed on the reward structure. This assumption is trivially satisfied in most applications.

ASSUMPTION 3.

$$M \overset{\text{def}}{=} \sup_{i \in I, a \in A(i)} r(i,a) < \infty. \tag{8}$$

To prepare the analysis in the following section we now introduce a data-transformation due to Schweitzer [26] which turns our semi-Markov decision model into an equivalent "pure" Markov decision model (MDP) with the same state space I, action sets $A(i), i \in I$ and with

$$\tilde{r}(i,a) = r(i,a)/\tau(i,a); \qquad i \in I, \quad a \in A(i); \tag{9}$$

$$\tilde{p}_{ij}(a) = \tau(p_{ij}(a) - \delta_{ij})/\tau(i,a) + \delta_{ij}; \qquad i,j \in I, \quad a \in A(i); \tag{10}$$

as the one-step expected rewards and transition probabilities. (δ_{ij} represents the Kronecker delta, i.e., $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.) The parameter τ is chosen such that $0 < \tau < \tau_{\min} \overset{\text{def}}{=} \inf_{i,a} \tau(i,a)/(1 - p_{ii}(a))$, (with $\tau_{\min} \geqslant \epsilon$ in view of Assumption 1), so as to ensure that $\tilde{p}_{ii}(a) > 0$ for all $i \in I, a \in A(i)$. (The nonnegativity of $\tilde{p}_{ij}(a)$, $i \neq j$, is automatically satisfied.) Note that for all $i \in I, \tilde{r}(i,\cdot)$ and $\tilde{p}_{ij}(\cdot)$, $j \in I$, are respectively upper semicontinuous and continuous on $A(i)$. Observe, in addition, that all stationary policies have *aperiodic* Markov chains in the transformed model since $\tilde{p}_{ii}(a) > 0$ for all $i \in I, a \in A(i)$. Finally, all quantities of interest in the transformed model will be denoted by a $\tilde{}$.

The following lemma proves inter alia that the average return optimality equation in the transformed model reads

$$v_i = \sup_{a \in A(i)} \left\{ \tilde{r}(i,a) - g + \sum_j \tilde{p}_{ij}(a) v_j \right\}, \qquad i \in I. \tag{11}$$

LEMMA 2. *Let Assumptions* 1–2 *hold.*

(a) *If* $\{g, v\}$ *is a solution to* (1) *with* $\sum_j p_{ij}(a)v_j$ *continuous on* $A(i)$, *for all* $i \in I$, *then* $\{g, \tau^{-1}v\}$ *is a solution to* (11); *conversely, if* $\{g, v\}$ *is a solution to* (11) *with* $\sum_j \tilde{p}_{ij}(a)v_j$ *continuous on* $A(i)$ *for all* $i \in I$, *then* $\{g, \tau v\}$ *is a solution to* (1).

(b) *For any measurable function* $h(i,a) \geqslant 0$, $i \in I$, $a \in A(i)$ *with*

$$\eta(i) \overset{\text{def}}{=} \sup_{f \in F} E_{f^\infty} \left\{ \sum_{k=0}^{N(K)-1} h(X_k, a_k) \mid X_0 = i \right\} < \infty, \qquad i \in I, \tag{12}$$

we have

$$\tilde{\eta}(i) \overset{\text{def}}{=} \sup_{f \in F} E_{f^\infty} \left\{ \sum_{k=0}^{\tilde{N}(K)-1} h(X_k, a_k)/\tau(X_k, a_k) \mid X_0 = i \right\} \leqslant \tau^{-1}\eta(i) + \gamma, \qquad i \in I,$$

for some constant $\gamma > 0$.

(c) *For all* $\psi \in \mathcal{P}, \psi^{(\infty)}$ *has the same sets of irreducible states both in the original and the transformed model. All of these sets are positive recurrent in both models; in particular*

(i) $K^* = \{ i \in K \mid i$ *is positive recurrent under some* $\tilde{P}(f), f \in F \}$,

(ii) *in both models,* $\psi^{*(\infty)}$ *has a single irreducible set of states which is positive recurrent and contains* K^*.

PROOF. (a) cf. the proof of Theorem 2.1 in Federgruen and Tijms [8].

(b) By results from positive dynamic programming (cf. Schäl [24] and [11, Lemma 2]):

$$\eta(i) = \max_{a \in A(i)} \left\{ h(i,a) + \sum_{j \notin K} p_{ij}(a)\eta(j) \right\}, \qquad i \in I,$$

or equivalently

$$0 \geqslant h(i,a) + \sum_{j \notin K} (p_{ij}(a) - \delta_{ij})\eta(j); \qquad i \notin K, \quad a \in A(i), \quad \text{and}$$

$$0 \geqslant h(i,a) + \sum_{j \notin K} (p_{ij}(a) - \delta_{ij})\eta(j) - \eta(i); \qquad i \in K, \quad a \in A(i).$$

Divide both inequalities by $\tau(i,a) > 0$ to obtain:

$$0 \geqslant \frac{h(i,a)}{\tau(i,a)} + \tau^{-1} \sum_{j \notin K} \left[\frac{\tau}{\tau(i,a)} (p_{ij}(a) - \delta_{ij}) + \delta_{ij} \right] \eta(j) - \tau^{-1}\eta(i); \qquad i \notin K, a \in A(i),$$

$$0 \geqslant \frac{h(i,a)}{\tau(i,a)} + \tau^{-1} \sum_{j \notin K} \left[\frac{\tau}{\tau(i,a)} (p_{ij}(a) - \delta_{ij}) + \delta_{ij} \right] \eta(j) - \frac{\eta(i)}{\tau(i,a)}, \qquad i \in K, a \in A(i).$$

Since K is finite and $\tau(i,a) \geqslant \epsilon$ for all (i,a) it follows that for some finite constant γ,

$$\tau^{-1}\eta(i) + \gamma\delta_{iK} \geqslant \frac{h(i,a)}{\tau(i,a)} + \tau^{-1} \sum_{j \notin K} \tilde{p}_{ij}(a)\eta(j), \qquad \text{for all} \quad i \in I, \quad a \in A(i) \quad (13)$$

where $\delta_{iK} = 1$ if $i \in K$ and 0 otherwise. By repeated iteration of this inequality and since $\eta \geqslant 0$ one concludes that for all $i \in I$ and $f \in F$ (cf. (3)):

$$\tau^{-1}\eta(i) + \gamma\delta_{iK} \geqslant \frac{h(i, f(i))}{\tau(i, f(i))} + \sum_{n=1}^{\infty} \sum_{j \notin K} \tilde{P}_{ij}^{n}(f) \frac{h(j, f(j))}{\tau(j, f(j))}$$

and hence $\tau^{-1}\eta(i) + \gamma \geqslant \tilde{\eta}(i)$, $i \in I$.

(c) Fix $\psi \in \mathcal{P}$. Since for $i \neq j$, $p_{ij}(a) > 0 \Leftrightarrow \tilde{p}_{ij}(a) > 0$, $a \in A(i)$ we have that C is an irreducible set in the original model $\Leftrightarrow C$ is an irreducible set in the transformed model.

Next fix an irreducible set C of $\psi^{(\infty)}$. In view of Lemma 1, C is positive recurrent in the original model. In view of (5) and [11, Lemma 2]

$$E_{\psi^{(\infty)}} \left\{ \sum_{k=0}^{N(K)-1} \tau(X_k, a_k) \,\middle|\, X_0 = i \right\} < \infty, \qquad i \in I.$$

In particular,

$$E_{\psi^{(\infty)}} \left\{ \sum_{k=0}^{N(K \cap C)-1} \tau(X_k, a_k) \,\middle|\, X_0 = i \right\} < \infty, \qquad i \in C.$$

To show that C is positive recurrent in the transformed model as well, consider an MDP with C as state space, and ψ as the single tpm, and apply [11, Theorem 2] to verify that $\exists i_0 \in K \cap C$ such that (12) holds with $K = \{i_0\}$, $h = \tau$ and $I = C$. Next apply part (b) of this lemma to conclude that $\tilde{\mu}_{ji}(\psi) < \infty$ for all $j \in C$. ∎

Choosing $h(i,a) = \tau(i,a)$ and $|r(i,a)|$, $i \in I$, $a \in A(i)$, respectively, in part (b) of the above lemma we have the existence of a constant $\gamma > 0$ such that

$$\begin{aligned}
&\sup_{f \in F} \tilde{\mu}_{iK}(f) \leqslant \tau^{-1}u^*(i) + \gamma, && i \in I, \\
&\sup_{f \in F} \tilde{\rho}_{iK}(f) \leqslant \tau^{-1}y^*(i) + \gamma, && i \in I, \quad \text{where} \\
&\tilde{\rho}_{iK}(f) = E_{f(\infty)} \left\{ \sum_{k=0}^{\tilde{N}(K)-1} |\tilde{r}(X_k, a_k)| \,\middle|\, X_0 = i \right\}, && i \in I, f \in F.
\end{aligned} \quad (14)$$

In the next section we prove the existence of a solution to the optimality equation (1) by exhibiting a convex compact subset of $E^\infty = X_{i\in I}E$, (with E the set of real numbers), which is mapped into itself by the value-iteration operator (cf. (22)). The boundary of this subset is defined by the numbers

$$z^*(i) \overset{\text{def}}{=} \sup_{f\in F} \tilde{\mu}_{iK^*}(f), \qquad i \in I \quad \text{and}$$

$$\rho_{ij}^* \overset{\text{def}}{=} E_{\psi^*(\infty)}\left\{ \sum_{k=0}^{\tilde{N}(\{j\})-1} \left(|\tilde{r}(X_k, a_k)| + \tilde{M}\right) \middle| X_0 = i \right\}, \qquad i \in I, \quad j \in K^*,$$

where

$$\tilde{M} \overset{\text{def}}{=} M\epsilon^{-1} \text{ is an upper bound on } \tilde{r}(i,a) = r(i,a)/\tau(i,a), \qquad i \in I, \quad a \in A(i). \tag{15}$$

We conclude this section by proving the finiteness of the numbers $\{z^*(i), \rho_{ij}^* \mid i \in I, j \in K^*\}$.

LEMMA 3. *Let Assumptions* 1–3 *hold. For any upper-semicontinuous and nonnegative function* $h(i,a)$, $i \in I$, $a \in A(i)$ *with* $\eta(i) < \infty$, $i \in I$ (*cf.* (12)) *we have for some constant* $c > 0$:

$$\text{(a)} \qquad \eta^*(i) = \sup_{a\in A(i)} \left\{ \frac{h(i,a)}{\tau(i,a)} + \sum_{j\notin K^*} \tilde{p}_{ij}(a)\eta^*(j) \right\} \leqslant \tau^{-1}\eta(i) + c < \infty, \qquad i \in I \tag{16}$$

where

$$\eta^*(i) \overset{\text{def}}{=} \sup_{f\in F} E_{f^\infty}\left\{ \sum_{k=0}^{\tilde{N}(K^*)-1} \frac{h(X_k, a_k)}{\tau(X_k, a_k)} \middle| X_0 = i \right\}.$$

$$\text{(b)} \qquad \eta_{ij}^* = \frac{h(i,\psi^*(i))}{\tau(i,\psi^*(i))} + \sum_{t\neq j} \tilde{p}_{it}(\psi^*(i))\eta_{tj}^* \leqslant \tau^{-1}\eta(i) + c, \qquad i \in I, \quad j \in K^*$$

where

$$\eta_{ij}^* \overset{\text{def}}{=} E_{\psi^*(\infty)}\left\{ \sum_{k=0}^{\tilde{N}(\{j\})-1} \frac{h(X_k, a_k)}{\tau(X_k, a_k)} \middle| X_0 = i \right\}; \qquad i \in I, \quad j \in K^*$$

and $h(i,\psi^*(i)) = E_{\psi^*(\infty)}h(i,a)$; $\tau(i,\psi^*(i)) = E_{\psi^*(\infty)}\tau(i,a)$ *and*

$$\tilde{p}_{ij}(\psi^*(i)) = E_{\psi^*(\infty)}\tilde{p}_{ij}(a); \qquad i,j \in I.$$

(c) *In particular there exists a constant* $c > 0$ *such that*

$$\begin{aligned} z^*(i) &= \sup_{a\in A(i)} \left\{ 1 + \sum_{j\notin K^*} \tilde{p}_{ij}(a)z^*(j) \right\} < \tau^{-1}u^*(i) + c < \infty, \qquad i \in I, \\ \rho_{ij}^* &= |\tilde{r}(i,\psi^*(i))| + \tilde{M} + \sum_{t\neq j} \tilde{p}_{it}(\psi^*(i))\rho_{tj}^* \\ &\leqslant \tau^{-1}\left(y^*(i) + \tilde{M}u^*(i)\right) + c, \qquad i \in I, \quad j \in K^*. \end{aligned} \tag{17}$$

PROOF. (a) We first prove the inequality for $\eta^*(i)$ in (16). The functional equation for $\eta^*(\cdot)$ then follows by applying a result from positive dynamic programming, see Schäl [25] and [11, Lemma 2]. Note that for $i \in K^*$, $\eta^*(i) \leqslant \tilde{\eta}(i) + \max_{j\in K\setminus K^*}\eta^*(j)$.

Hence it suffices to prove the existence of a constant $c > 0$ with

$$\eta^*(i) \leqslant \tau^{-1}\eta(i) + c < \infty, \qquad i \in I \backslash K^*. \tag{18}$$

Note that the value of $\eta^*(\cdot)$ on $I \backslash K^*$ remains unchanged when replacing the states in K^* by an aggregated state ∞, and when making this state absorbing under every policy in F. In this new model, let $h(i,a)/\tau(i,a)$ be the one-step expected reward when choosing action a in state i. Every policy in F has $\{\infty\}$ as its *single* irreducible set of states, cf. Lemma 1 part (a) and Lemma 2, part (c), so that Assumptions 1 and 2 in [11] are met. In view of [11, Theorem 2] there exists a number $c > 0$ such that for any $f \in F$ a state $s_f \in K$ exists for which

$$E_{f^{(\infty)}}\left\{ \sum_{k=0}^{\tilde{N}(\{s_f\})-1} \frac{h(X_k, a_k)}{\tau(X_k, a_k)} \,\middle|\, X_0 = i \right\} \leqslant \tau^{-1}\eta(i) + c, \qquad i \in I \backslash K^*.$$

It follows from relation (19) in [11] that s_f is recurrent under $f^{(\infty)}$. Hence, $s_f = \infty$ for all $f \in F$, which completes the proof of (18).

(b) Fix $j \in K^*$. Consider a MDP with $\hat{P}$ as the single Markov chain:

$$\hat{P}_{ik} = \begin{cases} \tilde{P}(\psi^*)_{ik}, & i \neq j, \\ \delta_{jk}, & k = j, \end{cases}$$

i.e., $\hat{P}$ has $\{j\}$ as its *single* irreducible set of states.

Let $h(i,\psi^*(i))/\tau(i,\psi^*(i))$ be the one-step expected reward in state i. The remainder of the proof is analogous to part (a).

(c) Part (c) follows from part (b), by the choices $h(i,a) = \tau(i,a)$ and $h(i,a) = |r(i,a)| + \tilde{M}\tau(i,a)$, $i \in I$ and $a \in A(i)$, as well as the definition of $u^*(i)$, $y^*(i)$, $i \in I$ (cf. Assumption 2). ∎

3. The average return optimality equation. In this section we prove the existence of a solution to the optimality equation (11) for the transformed model. In view of Lemma 2, part (a) this proves the existence of a solution to the optimality equation in the original model as well. We first define a bounding function; cf. Lippman [19]:

$$\mu^*(i) \overset{\text{def}}{=} \max_{j \in K^*} \rho^*_{ij} + z^*(i).$$

Note from Lemma 3 part (c) that a constant $c > 0$ exists such that

$$1 \leqslant \mu^*(i) \leqslant \tau^{-1}y^*(i) + \tau^{-1}(\tilde{M} + 1)u^*(i) + 2c, \qquad \text{all} \quad i \in I. \tag{19}$$

Let $E^\infty = X_{i \in I} E_i$, with E_i the real line, be endowed with the Euclidean product topology. For any $L > 0$, note that

$$V(L) \overset{\text{def}}{=} \{ x \in E^\infty \,|\, |x_i| \leqslant L\mu^*(i), \text{ for all } i \in I \} \tag{20}$$

is compact (in the product topology; see Tychonoff's theorem, Theorem 19 on p. 166 in Royden [23]). We finally need the following assumption, which is analogous to [11, Assumption 3].

ASSUMPTION 4. *For any* $i \in I$, $\sum_{j \in I} p_{ij}(a)\mu^*(j)$ *is continuous on* $A(i)$.

Using (10), and Assumption 1 one easily verifies

$$\text{for any } i \in I, \ \sum_{j \in I} \tilde{p}_{ij}(a)\mu^*(j) \text{ is continuous on } A(i). \tag{21}$$

Note in view of (19) that the finiteness $\sum_{j\in I}\tilde{p}_{ij}(a)\mu^*(j)$ for all $i \in I$, $a \in A(i)$ follows, since constants, c, γ_1, γ_2 exist such that

$$\begin{aligned}\sum_j \tilde{p}_{ij}(a)\mu^*(j) &\leqslant \sum_j \tilde{p}_{ij}(a)\left[\tau^{-1}y^*(j) + \tau^{-1}(\tilde{M}+1)u^*(j) + 2c\right] \\ &\leqslant \tau^{-1}y^*(i) + \gamma_2 + \left[\tau^{-1}(\tilde{M}+1)u^*(i) + (\tilde{M}+1)\gamma_1 + 2c\right] \\ &\quad + \sum_{j\in K}\tilde{p}_{ij}(a)\left[\tau^{-1}y^*(j) + \tau^{-1}(\tilde{M}+1)u^*(j)\right],\end{aligned}$$

the last inequality following from (13). Observe, in addition, that Assumption 4 is trivially satisfied in models with finite action spaces.

On $V^* \overset{\text{def}}{=} U_{r>0}V(r)$, we define the following two (value iteration) operators:

$$Tx_i = \sup_{a\in A(i)}\left\{\tilde{r}(i,a) + \sum_j \tilde{p}_{ij}(a)x_j\right\}, \qquad i \in I; \quad \text{and} \tag{22}$$

$$Qx_i = Tx_i - Tx_{i^*}, \qquad i \in I, \quad \text{where } i^* \text{ is some fixed state in } K^*. \tag{23}$$

Note that for $x \in V^*$,

$$\sum_j \tilde{p}_{ij}(a)|x_j| \leqslant \left\{\sup_{j\in I}\mu^*(j)^{-1}|x_j|\right\}\sum_j \tilde{p}_{ij}(a)\mu^*(j) < \infty$$

so that the expressions between { } in (22) are well defined. In addition, for any $i \in I$, $\sum_j \tilde{p}_{ij}(a)x_j$ is continuous in $a \in A(i)$ in view of Assumption 4, $|x_j| \leqslant \{\sup_{l\in I}\mu^*(l)^{-1}|x_l|\}\,\mu^*(j), j \in I$ and Royden [23, Proposition 18, p. 232]. This, together with the upper-semicontinuity of $\tilde{r}(i,\cdot)$ on $A(i)$, $i \in I$ and the compactness of the sets $A(i)$, $i \in I$ (cf. Assumption 2) imply that the supremum in (22) may be replaced by a maximum; see Royden [23, Proposition 10, p. 161].

Our analysis is based on the construction of a compact, convex subset of E^∞ which is closed for the Q-operator. Define $R^* = \max_{i,j\in K^*}\rho_{ij}^*$ and recall (15) for the definition of $\tilde{M}$. Define (cf. Lemma 3 and Assumptions 1–3):

$$D_1 = \{x \mid x_i - x_j \geqslant -\rho_{ij}^*, i \in I, j \in K^*\}, \tag{24}$$

$$D_2 = \{x \in V^* \mid Tx_i \leqslant x_i + \tilde{M}, i \in I\}, \tag{25}$$

$$D_3 = \{x \mid x_i - x_j \leqslant (\tilde{M} + 2R^*)z^*(i), i \in I\backslash K^*, j \in K^*\}, \tag{26}$$

$$D_4 = \{x \mid x_{i^*} = 0\}.$$

Finally, let $D = D_1 \cap D_2 \cap D_3 \cap D_4$.

THEOREM 1. *Let Assumptions 1–3 hold.*
(a) *T maps $D_1 \cap D_2 \cap D_3$ into itself;*
(b) *D is a nonempty, convex compact subset of $V(\tilde{M} + 2R^*)$;*
(c) *Q is a continuous operator on D and maps D into itself.*

PROOF. (a) (I) if $x \in D_1 \cap D_2 \cap D_3$, show $Tx \in D_1$ as follows:
Fix $j \in K^*$ and $i \in I\backslash\{j\}$. Note that

$$Tx_i \geqslant \tilde{r}(i,\psi^*(i)) + \tilde{p}_{ij}(\psi^*(i))x_j + \sum_{t\neq j}\tilde{p}_{it}(\psi^*(i))x_t.$$

Insert $x_t \geqslant x_j - \rho_{tj}^*$ for $t \neq j$ and $x_j \geqslant Tx_j - \tilde{M}$ (in view of $x \in D_1 \cap D_2$) and conclude

using (17):

$$Tx_i \geqslant \tilde{r}(i,\psi^*(i)) + x_j - \sum_{t\neq j} \tilde{p}_{it}(\psi^*(i))\rho^*_{tj}$$

$$\geqslant -|\tilde{r}(i,\psi^*(i))| - \tilde{M} + Tx_j - \sum_{t\neq j} \tilde{p}_{it}(\psi^*(i))\rho^*_{tj} = -\rho^*_{ij} + Tx_j .$$

(II) If $x \in D_1 \cap D_2$ show $Tx \in D_2$ as follows: Let $1 \in E^\infty$ be the vector all of whose components are unity. It follows from the monotonicity of the T-operator that $T^2x \leqslant T(x + \tilde{M}1) = Tx + \tilde{M}1$ thus showing that Tx satisfies the inequalities in the definition of D_2. We now show $Tx \in V^*$. Let $x \in V(c)$ for some $c > 0$. In view of (25), $Tx_i/\mu^*(i) \leqslant x_i/\mu^*(i) + \tilde{M}/\mu^*(i) \leqslant c + \tilde{M}$ since $\mu^*(i) \geqslant 1$, $i \in I$. To establish that $\{Tx_i/\mu^*(i), i \in I\}$ is uniformly bounded from below as well, note in view of (24) that $x_i \geqslant -\rho^*_{ii^*} + x_{i^*}$. Hence for all $i \in I$, and $a \in A(i)$, $Tx_i \geqslant \tilde{r}(i,a) - \sum_j \tilde{p}_{ij}(a)\rho^*_{ji^*} + x_{i^*}$. This implies

$$Tx_i \geqslant \tilde{r}(i,\psi^*(i)) - \sum_j \tilde{p}_{ij}(\psi^*(i))\rho^*_{ji^*} + x_{i^*}$$

$$\geqslant -\left\{|\tilde{r}(i,\psi^*(i))| + \sum_j \tilde{p}_{ij}(\psi^*(i))\rho^*_{ji^*}\right\} + x_{i^*}, \qquad i \in I. \tag{27}$$

In view of (17) there exists a constant M_1 such that

$$\rho^*_{ii^*} = |\tilde{r}(i,\psi^*(i))| + \tilde{M} + \sum_{j\neq i^*} \tilde{p}_{ij}(\psi^*(i))\rho^*_{ji^*}$$

$$\geqslant |\tilde{r}(i,\psi^*(i))| + \sum_{j\in I} \tilde{p}_{ij}(\psi^*(i))\rho^*_{ji^*} + M_1 .$$

Insert this inequality into (27) to conclude for all $i \in I$ that $Tx_i \geqslant -\rho^*_{ii^*} + M_1 + x_{i^*}$ and hence in view of $\mu^*(i) \geqslant 1$, $i \in I$, $Tx_i/\mu^*(i) \geqslant -(+1 + |x_{i^*}| + |M_1|)$, $i \in I$.

(III) If $x \in D_1 \cap D_2 \cap D_3$, show $Tx \in D_3$ as follows: Fix $i \in I \setminus K^*$ and $j \in K^*$. Note that

$$Tx_i = \max_{a\in A(i)} \left\{\tilde{r}(i,a) + \sum_{t\in K^*} \tilde{p}_{it}(a)x_t + \sum_{t\notin K^*} \tilde{p}_{it}(a)x_t\right\}. \tag{28}$$

Let a^* achieve the maximum in (28) and insert for $t \notin K^*$, in view of $x \in D_3$, $x_t \leqslant (\tilde{M} + 2R^*)z^*(t) + [\max_{i\in K^*} x_i]$

$$Tx_i \leqslant \tilde{r}(i,a^*) + \left[\max_{i\in K^*} x_i\right] + (\tilde{M} + 2R^*) \sum_{t\notin K^*} \tilde{p}_{it}(a^*)z^*(t). \tag{29}$$

Next let $l \in K^*$ and conclude from repeated substitutions of the inequality $Tx_j \geqslant \tilde{r}(j,\psi^*(j)) - \tilde{M} + \sum_t \tilde{p}_{jt}(\psi^*(j))Tx_t$ that

$$Tx_j \geqslant -E_{\psi^*(\infty)}\left[\sum_{k=0}^{\tilde{N}(\{l\})-1} (|\tilde{r}(X_k,a_k)| + \tilde{M}) \,\middle|\, X_0 = j\right] + x_l$$

$$= -\rho^*_{jl} + x_l \geqslant -R^* + \left[\min_{l\in K^*} x_l\right].$$

Finally, subtract this inequality from (29) to conclude, using $\max_{l\in K^*} x_l - \min_{l\in K^*} x_l \leqslant R^*$ (in view of $x \in D_1$):

$$Tx_i - Tx_j \leqslant (\tilde{M} + 2R^*) + (\tilde{M} + 2R^*) \sum_{t\notin K^*} \tilde{p}_{it}(a^*)z^*(t)$$

$$\leqslant (\tilde{M} + 2R^*) \max_{a\in A(i)} \left\{1 + \sum_{t\notin K^*} \tilde{p}_{it}(a)z^*(t)\right\} = (\tilde{M} + 2R^*)z^*(i).$$

(b) $0 \in D$, hence $D \neq \emptyset$. We next show that

$$D_1 \cap D_3 \cap D_4 \subseteq V(\tilde{M} + 2R^*), \tag{30}$$

thus showing that D is a subset of a compact space. From (24) with $j = i^*$ we obtain $-\rho^*_{ii^*} \leqslant x_i$ for all $i \in I$, and from (26) with $j = i^*$ we get: $x_i \leqslant (\tilde{M} + 2R^*)z^*(i)$, $i \notin K^*$ whereas for $i \in K^*$ the upperbound $x_i \leqslant \rho^*_{i^*i} \leqslant R^*\mu^*(i)$ follows from (24) with $i = i^*$ and $j = i$.

The upper and lower bounds on the components of x establish (30), cf. (19). Hence to prove compactness of D, we merely have to show its closedness. Take a sequence $\{x^{(n)}\}_{n=1}^{\infty}$ in D with $\lim_{n\to\infty} x^{(n)} = x^*$. One immediately verifies $x^* \in D_1 \cap D_3 \cap D_4$ and hence $x^* \in V(\tilde{M} + 2R^*)$ in view of (30). To show $x^* \in D_2$ we first need

$$\lim_{n\to\infty} Tx_i^{(n)} = \lim_{n\to\infty} \sup_{a\in A(i)} \left\{ \tilde{r}(i,a) + \sum_j \tilde{p}_{ij}(a) x_j^{(n)} \right\} = Tx_i^*, \qquad i \in I. \tag{31}$$

(Note that the functions within { } are upper-semicontinuous on $A(i)$ for all $n \geqslant 1$, use (21), $x^{(n)} \in V(\tilde{M} + 2R^*)$, Royden [23, Proposition 18, p. 232] and Schäl [25, Proposition 10.1]).

Since convergence in D_T is pointwise, we obtain for any $i \in I$, $x_i^* + \tilde{M} \geqslant \lim_{n\to\infty} x_i^{(n)} + \tilde{M} \geqslant \lim_{n\to\infty} Tx_i^{(n)} = Tx_i^*$, hence $x^* \in D_2$. Finally, the convexity of D follows from the convexity of D_i, $i = 1, \ldots, 4$: D_1, D_3, D_4 are polyhedral sets, and D_2 is convex since V^* is convex, and in view of the inequality

$$T(\lambda x + (1-\lambda)y) \leqslant \lambda Tx + (1-\lambda)Ty, \qquad 0 \leqslant \lambda \leqslant 1.$$

(c) We fix $x \in D$ and first show $Qx \in D$. (I) $Qx \in D_1$, since for all $j \in K^*$, $i \neq j$: $Qx_i - Qx_j = Tx_i - Tx_j \geqslant -\rho^*_{ij}$ in view of part (a).

(II) $Qx \in D_2$ since $T(Qx) = T(Tx - (Tx)_{i^*}1) = T^2x - (Tx)_{i^*}1$. Insert $T^2x \leqslant Tx + \tilde{M}1$ (in view of $Tx \in D_2$, cf. part (a)):

$$T(Qx) \leqslant Tx - (Tx)_{i^*}1 + \tilde{M}1 = Qx + \tilde{M}1, \quad \text{so} \quad Qx \in D_2.$$

(III) $Qx \in D_3$ since for all $i \notin K^*$, $j \in K^*$:

$$Qx_i - Qx_j = Tx_i - Tx_j \leqslant (\tilde{M} + 2R^*)z^*(i)$$

in view of part (a). (IV) $Qx \in D_4$ since $(Qx)_{i^*} = 0$. The continuity of the Q-operator follows immediately from the continuity of the T-operator. Since convergence in E^∞ is pointwise, the latter results from (31). ∎

In view of Theorem 1, an extension of Brouwer's fixed point theorem to infinite dimensional vector spaces establishes the existence of a fixed point of the operator Q on D and hence the existence of a solution to the optimality equation.

THEOREM 2 (MAIN THEOREM). *Let Assumptions* 1–4 *hold.*

(a) *The Q-operator has a fixed point* $\{v_i \mid i \in I\}$ *on* D.

(b) *There exists a constant* g^* *and a vector* $\{v_i, i \in I\}$ *with* $\sup_{j\in I}|v_j|/\mu^*(j) < \infty$ *satisfying the optimality equation* (11).

PROOF. (a) We note that E^∞ is a locally convex linear topological space. By parts (b) and (c) of the previous theorem D is a compact convex subset of E^∞ and Q is a continuous operator mapping D into itself. Invoke Tychonoff's fixed point theorem (cf. Dugundji [7, Theorem 2.2, p. 414]). (b) Use part (a) with $g^* \overset{\text{def}}{=} (Tv^*)_{i^*}$. ∎

In view of Lemma 2, we conclude that under Assumptions 1–4, the optimality equation (1) of the original model has a solution as well. Existence of a solution to the optimality equation (11) is in itself *insufficient* for the existence of an optimal stationary policy as is exhibited by the example in Fisher and Ross [15] which satisfies

our Assumptions 1–4 and where *no* stationary policy dominates within the class of all stationary policies. Only in case the optimality equation (1) has a solution which is bounded in the L_∞-norm, is optimality guaranteed of any policy which, for every i, prescribes an action maximizing the right-hand side of (1). Hence an additional assumption is required (cf. also Robinson [20] and [21]). From here we follow the analysis in [11] by imposing an analogous version of Assumption 4 ibid. First for any $f \in F$, define the substochastic matrix $\hat{P}(f)$ as a truncation of the matrix $P(f)$ in the *original* model.

$$\hat{P}(f)_{ij} = \begin{cases} P(f)_{ij} & \text{for } i \in I, \quad j \notin K, \\ 0 & \text{for } i \in I, \quad j \in K. \end{cases}$$

ASSUMPTION 5. *For any $f \in F$, $\lim_{n\to\infty} \hat{P}^n(f)\mu^* = 0$ where $\hat{P}^n(f)$ denotes the n-fold matrix product of the substochastic matrix $\hat{P}(f)$ with itself.*

(In view of (19) Assumption 5 may be verified when replacing the function μ^* by $u^* + y^*$.) The following theorem follows from the proof of Theorem 5 in [11].

THEOREM 3. *Suppose that Assumptions 1–5 hold. A solution $\{g; v_i \mid i \in I\}$ to the optimality equation (1) exists. Choose any stationary policy $f^{(\infty)}$ such that the action $f(i)$ maximizes the right side of (1) for this solution and all $i \in I$. Then policy $f^{(\infty)}$ is average optimal (in the strong sense) and g is uniquely determined as the maximal gain rate.*

Observe that whereas g is uniquely determined by (1) (under the additional Assumption 5), the v-function never is: note that if (g, v) is a solution to (1), then so is $(g, v + c1)$ for any scalar c. [11] showed for *unichained* models that the solution to the optimality equation is in fact unique up to this additive constant for the v-vector, under the following strengthening of Assumption 5.

ASSUMPTION 5′. $\sup_{f\in F} E_{f^{(\infty)}}\{\sum_{k=0}^{N(K)-1} \mu^*(X_k) \mid X_0 = i\} < \infty$.

(Assumption 5′ requires the supremum over all stationary policies of the total reward until the first visit to K to be finite for every starting state, given that $\mu^*(i)$ represents the one step reward in state i. Again, Assumtion 5′ may be verified by replacing the "one-step reward" function μ^* by $u^* + y^*$, cf. (19).)

In our models where (even optimal) policies may have *multiple* irreducible sets of states, the solution space is more complex as follows from the following 2-state example (cf. also [27] where it is shown that the solution set may even be nonconvex).

EXAMPLE 2. $I = \{1, 2\}$; $A(i) = \{0, 1\}$, $i = 1, 2$.

i	a	$p_{11}(a)$	$p_{i2}(a)$	$r(i,a)$
1	0	1	0	0
	1	0	1	−1
2	0	0	1	0
	1	1	0	−1

states 1 and 2 reach each other: $\{(g, v) \mid g = 0 \text{ and } |v_1 - v_2| \leqslant 1\}$ is the solution set to (1).

The following theorem provides a partial characterization of the solution set of the optimality equation (11). (We refer to [14] for a proof.) In particular it points out that the solution set of (1) is bounded in the quasi-norm $\mathrm{sp}[x] = \sup_i x_i/\mu^*(i) - \inf_i x_i/\mu^*(i)$. This represents a direct generalization of the results in Bather [1].

THEOREM 4. *Let Assumptions 1–4, 5′ hold. Every solution $\{g, v_i \mid i \in I\}$ of (11) with $v \in V^*$ satisfies $v \in D_1 \cap D_2 \cap D_3$. Moreover $v^* = v - (v_{i^*})1 \in D$ and v^* is a fixed point of Q.*

Acknowledgement. The authors thank Mordecai Haimovich for helpful suggestions. In addition, they are indebted to the referees for many useful comments.

Appendix 1. In this appendix we verify Assumptions 1–4 for the model described in Example 1. To simplify the notation, we take $n = 2$ and assume that the $h(\cdot)$ function is bounded and as in [30] and [31], that the probabilities $\{p_i, i > 0\}$ are nonincreasing in $i \geqslant 0$. As a slight generalization of [31] we assume $p_i < 1$ for $i \geqslant L_2 \geqslant 0$. In fact the conditions in this paper for the existence of a stationary optimal policy can be verified for a general nondecreasing $h(i)$ function which is bounded by some polynomial in i.

$I = R^2$ where $R = \{-L_1, -L_1 + 1, \ldots\}$; here $i \in R$ indicates for $i \geqslant 0$, the number of periods since the last overhaul and for $-L_1 \leqslant i \leqslant -1$, $|i|$ equals the number of periods to go until the completion of the current overhaul. $A(i, j) = B(i) \times B(j)$ for all $i, j \in 1$ where $B(i) = \{0, 1\}$ if $i \geqslant 0$ and $B(i) = \{0\}$ if $i < 0$. (Here the alternative $a = 1$ [0] represents a decision to [not to] overhaul.) If alternative 1 is chosen for some component, its age in the next period is $-L_1 + 1$; if a component fails, its age in the next period is $-L_1$. In view of the finiteness of the action sets, Assumptions 1 and 4 are trivially met; since all costs are nonnegative, Assumption 3 is met as well. To verify Assumption 2, we choose $K = \{-L_1, \ldots, L_2\}^2$. Note that for $(i, j) \in I \setminus K$, there is under any policy a probability of at least $(1 - p_{L_2})^2$ of being in K at the end of the first period, either because of failure or a decision to overhaul one or both of them. This implies that the simultaneous Doeblin condition is satisfied with respect to K, and hence (5) and (6) are satisfied (cf. Hordijk [17, Theorem 11.3], or [10, Theorem 2.2]), with $\sup_i u^*(i) < \infty$, $\sup_i y^*(i) < \infty$. Hence $\sup_i \mu^*(i) < \infty$, cf. (17).

Next, note that all states in K reach each other under the policy which *never* overhauls. Thus, the optimality equation (1) has a solution which is *bounded* since $\sup_i \mu^*(i) < \infty$. The existence of a stationary optimal policy then follows from Ross [22, Theorem 7.6]. (Note however, as in Stengos [30, p. 85] that multichain policies exist. Consider, e.g., the extremely conservative policy which overhauls as soon as a piece of equipment reaches age L_2, which may even be optimal for some $h(\cdot)$ functions. This policy has $L \overset{\text{def}}{=} L_1 + L_2$ subchains C_k, $k = 0, \ldots, L - 1$ where $C_k = \{(i, j) \mid -L_1 \leqslant i, j \leqslant L_2 - 1$ and $i - j = k$ (modulo $L)\}$.)

Appendix 2. In a single item system in [9], [28] and [29] demands occur at epochs generated by a Poisson process with rate λ. The demand sizes are independent positive random variables with common discrete probability distribution $\{\phi(j), j \geqslant 0\}$. There are two types of ordering opportunities. "Normal" (high-cost) opportunities occur at the demand epochs whereas special opportunities (at low cost) occur at epochs generated by a Poisson process with rate μ, assumed to be independent of the demand process.

Demand epochs and "special" replenishment opportunities representing the decision epochs, the state space is given by $X = \{(i, z) \mid i$ integer; $z = 0, 1\}$. Here state $(i, 0)[(i, 1)]$ corresponds to the situation where a demand [special replenishment opportunity] has just occurred leaving a net inventory position (= (inventory on hand) + (outstanding orders) − (backlog)) of i units. At each decision epoch we specify the decision l as the inventory position just after a possible replenishment. There are holding costs at a nonnegative rate $h(i)$ when the inventory on hand equals i, and a rate $h(-i)$ for a backlog of i units. We assume $h(i) = O(|i|^m)$ for some $m > 0$.

Without loss of generality we assume the existence of integers $L < 0$ and $U > 0$ such that the action sets $A(x)$ can be restricted to $A(i, 0) = A(i, 1) = \{l \mid \max(i, L) \leqslant l \leqslant U;$ $r = 0, 1\}$ for $i \leqslant U$ and $A(i, 0) = A(i, 1) = \{i\}$ for $i > U$. Assumptions 1, 3 and 4 are trivially met since all $A(x)$ are finite and the one-step rewards $r(\cdot, \cdot) \leqslant 0$. To prove the

existence of a solution to the optimality equation, it thus suffices to verify Assumption 2. Let $K = \{(i,r) \mid L \leqslant i \leqslant U;\ r = 0,1\}$. Note from [9, (2.2)] that $r(x,l) = O(|l|^{m+1})$. Part (a) of Assumption 2 being trivial for starting states $x = (i,r)$ with $i \leqslant U$, we consider the case $i > U$. Since replenishments are avoided as long as the net inventory position exceeds U, the expected time until reaching the set K is bounded from above by $(\lambda\phi(0))^{-1}(i - U) < \infty$ and the expected absolute reward by an expression which is $\sum_{l=U+1}^{i}(\lambda\phi(0))^{-1}O(|l|^{m+1}) = O(|l|^{m+2}) < \infty$. Hence

$$\mu^*(i,r) = O(|i|^{m+2}); \qquad \text{all } i,r. \tag{A1}$$

To verify part (b) of Assumption 2, fix $y^0 = (i,r) \in K^*$, let $f^0 \in F$ be a policy under which y^0 is recurrent and let C be the ergodic subchain (under $f^{0(\infty)}$) which contains y^0. Let $j^* = \min\{j \geqslant 0 \mid \phi(j) > 0\}$ and note, that if $y^0 = (i,0)$, $i \leqslant U - j^*$. This follows from the fact that in every recurrent predecessor state (i^1, r^1) of y^0 (under $f^{0(\infty)}$) we either have $i^1 \geqslant i + j^*$ and $i^1 \leqslant U$ or a replenishment is prescribed which increases the net inventory position to a level $l \leqslant U$ and enables transitions at the next decision epoch to states $(i,0)$ with $0 \leqslant l - j^* \leqslant U - j^*$ only. Finally fix $x^0 \in K$ and let $S(x^0)$ be the set of states in K that can be reached from x^0 when avoiding replenishments. If $S(x^0) \cap C \neq \emptyset$, it is easy to construct a rule f^1 under which y^0 can be reached from x^0. Otherwise select a state $(i^*, r^*) \in S(x^0)$ with $i^* < i$ and construct f^1 as follows:

$$f^1(x) = \begin{cases} f^0(x), & x \in C, \\ i + j^* \leqslant U, & \text{for } x = (i^*, r^*) \quad \text{if } y^0 = (i,0), \\ i, & \text{for } x = (i^*, r^*) \quad \text{if } y^0 = (i,1), \\ \max(L, j), & \text{for all other } x = (j, \cdot). \end{cases}$$

Note that y^0 can be reached from (i^*, r^*) and $(i^*, r^*) \in S(x^0)$ can be reached from x^0 under f^1.

Finally, to prove the existence of an optimal stationary strategy, one verifies Assumption (5) via its stronger version (5′). The verification of Assumption 5′ is identical to the verification of Assumption 2(a) using (A1). Note that some of the policies have multiple subchains.

References

[1] Bather, J. (1973). Optimal Decision Procedures for Finite Markov Chains, Part II. *Adv. in Appl. Probab.* **5** 521–540.

[2] Chung, K. (1967). Markov Chains with Stationary Transition Probabilities (2nd ed.). Springer, Berlin.

[3] Deppe, H. (1981). On the Existence of Average Cost Optimal Policies in Semi-regenerative Decision Models. Preprint 476, Institute for Applied Mathematics, University of Bonn.

[4] Derman, C. (1966). Denumerable State Markovian Decision Processes—Average Cost Criterion. *Ann. Math. Statist.* **37** 1545–1553.

[5] ——— and Strauch, R. (1966). A Note on Memoryless Rules for Controlling Sequential Processes. *Ann. Math. Statist.* **37** 276–278.

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

[7] Dugundji, J. (1970). *Topology*. Allyn and Bacon, Boston. (1980) 5th ed.

[→] Federgruen, A. and Tijms, H. C. (1978). The Optimality Equation in Average Cost Denumerable State Semi-Markov Decision Processes. Recurrency Conditions and Algorithms. *J. Appl. Probab.* **15** 356–373.

[9] ———, Groenevelt, H. and Tijms, H. C. (1982). Coordinated Replenishments in a Multi-item Inventory System with Compound Poisson Demands and Constant Lead Times. Columbia University Graduate School of Business, Working Paper No. 362A, New York.

[10] ———, Hordijk, A. and Tijms, H. C. (1978). Recurrence Conditions in Denumerable State Markov Decision Process. In *Dynamic Programming and Its Applications*, M. L. Puterman, ed. Academic Press, New York.

[11] Federgruen, A., Hordijk, A. and Tijms, H. C. (1979). Denumerable State Semi-Markov Decision Processes with Unbounded Costs, Average Cost Criterion. *Stochastic Process. Appl.* **9** 222–235.
[12] ———, ——— and ———. (1978). A Note on Simultaneous Recurrence Conditions on a Set of Denumerable Stochastic Matrices. *J. Appl. Probab.* **15** 842–847.
[13] ——— and Schweitzer, P. J. (1980). A Fixed Point Approach for Undiscounted Markov Renewal Programs. Columbia University Graduate School of Business, Working Paper No. 351A.
[14] ———, Schweitzer, P. J. and Tijms, H. C. (1981). Denumerable Undiscounted Semi-Markov Decision Processes with Unbounded Rewards. Columbia University Graduate School of Business, Working Paper No. 355A, New York (unabridged version of this paper).
[15] Fisher, L. and Ross, S. M. (1968). An Example in Denumerable Decision Processes. *Ann. Math. Statist.* **39** 674–675.
[16] Flynn, J. (1976). Conditions for the Equivalence of Optimality Criteria in Dynamic Programming. *Ann. Statist.* **4** 936–953.
[17] Hordijk, A. (1974). Dynamic Programming and Markov Potential Theory. Mathematical Centre Tract No. 51, Mathematisch Centrum, Amsterdam.
[18] ———. (1976). Regenerative Markov Decision Models. In *Mathematical Programming Study* **6** R. Wets, ed. North-Holland, Amsterdam, 49–72.
[19] Lippman, S. (1975). On Dynamic Programming with Unbounded Rewards. *Management Sci.* **21** 348–357.
[20] Robinson, D. (1976). Markov Decision Chains with Unbounded Costs and Applications to the Control of Queues. *Adv. in Appl. Probab.* **8** 159–176.
[21] ———. (1980). Optimality Conditions for a Markov Decision Chain with Unbounded Costs. *J. Appl. Probab.* **17** 996–1003.
[22] Ross, S. M. (1970). *Applied Probability Models with Optimization Applications*. Holden-Day, San Francisco.
[23] Royden, H. L. (1968). *Real Analysis*. Macmillan, New York, 2nd ed.
[24] Schäl, M. (1971). Ein Veralgemeinteres Stationäres Entscheidungs-modell der Dynamischen Optimierung. In *Methods of Operations Research* **10** 145–162.
[25] ———. (1975). Conditions for Optimality in Dynamic Programming and for the Limit of n-Stage Optimal Policies to be Optimal. *Z. Wahrsch. Verw. Gebiete* **32** 179–196.
[26] Schweitzer, P. J. (1971). Iterative Solution of the Functional Equations for Undiscounted Markov Renewal Programming. *J. Math. Anal. Appl.* **34** 495–501.
[27] ——— and Federgruen, A. (1978). Functional Equations of Undiscounted Markov Renewal Programming. *Math. Oper. Res.* **3** 308–322.
[28] Silver, E. (1974). A Control System for Coordinated Inventory Replenishment. *Internat. J. Prod. Res.* **12** 647–671.
[29] ——— and Thompstone, R. (1975). A Coordinated Inventory Control System for Compound Poisson Demand and Zero Lead Time. *Internat. J. Prod. Res.* **13** 581–602.
[30] Stengos, D. *Finite State Approximations for Denumerable State Markov Decision Processes—The Average Cost Case* (to appear).
[31] ——— and Thomas, L. (1980). The Blast Furnaces Problem. *European J. Oper. Res.* **4** 330–337.
[32] Taylor, H. (1965). Markovian Sequential Replacement Processes. *Ann. Math. Statist.* **36** 1677–1694.
[33] Wijngaard, J. (1977). Stationary Markov Decision Problems and Perturbation Theory of Quasi-Compact Linear Operators. *Math. Oper. Res.* **2** 91–102.

FEDERGRUEN: GRADUATE SCHOOL OF BUSINESS, COLUMBIA UNIVERSITY, NEW YORK, NEW YORK 10027
SCHWEITZER: GRADUATE SCHOOL OF MANAGEMENT, UNIVERSITY OF ROCHESTER, ROCHESTER, NEW YORK 14627
TIJMS: DEPARTMENT OF ACTUARIAL SCIENCES AND ECONOMETRICS, VRIJE UNIVERSITEIT, AMSTERDAM, THE NETHERLANDS