æ

Contraction conditions for average and α-discount optimality in countable state Markov games with unbounded rewards

E. Altman[1), A. Hordijk[2) and F.M. Spieksma[2) 3)

June, 1996

ABSTRACT.

The goal of this paper is to provide a theory of N-person Markov games with unbounded cost, for a countable state space and compact action spaces. We investigate both the finite and infinite horizon problems. For the latter, we consider the discounted cost as well as the expected average cost. We present conditions for the infinite horizon problems for which equilibrium policies exist for all players within the stationary policies, and show that the costs in equilibrium satisfy the optimality equations. Similar results are obtained for the finite horizon costs, for which equilibrium policies are shown to exist for all players within the Markov policies. As special case of N-person games, we investigate the zero-sum (2 players) game, for which we establish the convergence of the value iteration algorithm. We conclude by studying an application of a zero-sum Markov game in a queueing model.

KEYWORDS.

Non-cooperative Markov games, μ-geometric recurrence, equilibrium policies, value iteration, birth-death control.

[1)] Postal address: INRIA Sophia Antipolis, B.P.93, 2004 Rue des Lucioles, 06902 Sophia Antipolis Cedex, France

[2)] Postal address: Dept. of Mathematics and Computer Science, Leiden University, P.O.Box 9512, 2300RA Leiden, The Netherlands

[3)] The research of this author has been supported by the Netherlands Organisation for Scientific Research N.W.O. and by a fellowship from the Royal Netherlands Academy of Arts and Sciences

1. Introduction.

In this article we study non-cooperative N-person Markov games on a countable state space E with compact action spaces, such that the immediate rewards associated with each state-action pair are unbounded in the state variable. The fact that the state space is infinite and the rewards are unbounded, requires special attention. Standard statements that are typically obtained using dynamic programming no longer hold. Indeed, even when there is a single player, the dynamic programming equations frequently have infinitely many solutions (see e.g. [2] Example 4.1). This may be the case both for the discounted as well as the expected average cost criteria. In other cases, the dynamic programming equations may have no solutions at all.

Our analysis is based on the use of contraction conditions for a normed Banach space. This approach will ensure existence (and uniqueness in the case of zero-sum games) of solutions to the dynamic programming equations in this normed space. More precisely, we make the following contraction assumption on the Markov chain: there exists a constant $\beta < 1$ and a finite set $M \subset \mathsf{E}$, such that for any stationary policy π of the players

$$\alpha \sup_{i \in \mathsf{E}} \mu_i^{-1} \sum_{j \notin M} P_{ij}(\pi) \mu_j \leq \beta, \tag{1.1}$$

where $\alpha \in (0, 1]$ is the discount factor, $\mu \geq 1$ a given positive vector on E (cf. §2). For $\alpha = 1$ this property is called μ-uniform geometric recurrence (cf. [11, 12]).

If the immediate rewards are bounded by μ (cf. §2), then we will show in §2 for the N-person non-cooperative game that there is a stationary equilibrium policy for the infinite horizon α-discounted optimality criterion, i.e. there is a stationary policy such that no player will gain by deviating, provided some suitable continuity conditions hold. For the finite horizon case we show that there is a Markov equilibrium policy for all players.

The average reward case then follows under condition (1.1) for $\alpha = 1$ from well-known limiting arguments. These results extend in a straightforward manner (by showing that an average equilibrium policy is a *strong* average equilibrium point) to 2-person zero-sum games. For the latter we will show in §3 under an additional strong unichain assumption, that the value iteration algorithm for average rewards converges.

As a motivation for our analysis, we study in §4 a time-discretised control of a birth-death process, where the death rates are controlled, but the birth rates are only known to assume values in some interval (depending on the actual state of the process). An application of this model is service control in a queueing network consisting of one service centre, whilst the arrival rate is unknown. We model this situation as a zero-sum game against "nature" (cf. [1, 3, 5]). For flexibility reasons it is to preferred not to put some pre-determined bound on the feasible queue size. This motivates the countable state space assumption.

Additionally, knowing the structure of an equilibrium policy will considerably reduce computational efforts. Because of the lack of boundary effects, the problem of determining such structures is easier for the countable state system. Therefore the countable state system can be considered as a convenient approximation of finite state (but large) systems, because the boundary will only influence the policy in the rarely visited part of the state space.

However, under the countable state assumption it is then necessary to allow unbounded rewards, since in queueing applications the immediate costs/rewards are typically polynomial functions of the queue length. For deriving structures of equilibrium polices the value-iteration algorithm is convenient, thus motivating the need for this algorithm to converge even for non-computer applications.

In §4 we will model the controlled birth-death process as a 2-person zero-sum game. A similar model was analysed in [3]. However it had several restrictions. To start with, [3] only considered discounted costs, as the theory for the expected average costs case was lacking. Secondly it required the state process to be ergodic under any stationary multi-strategy of both players. In this paper under essentially condition (1.1) we will derive the existence as well as the structure of an equilibrium policy. However (1.1) is a heavy condition in that it requires ergodicity of the process controlled by any stationary policy. In section 4 we will show that this condition can be relaxed to allow policies under which the state process becomes transient: our arguments (cf. [14, 22]) are based on the convexity of the value function and the left skipfree property of the transition kernels.

Next, we mention and discuss some related work on unbounded costs. Condition (1.1) is equal to the weighted supremum norm condition used by Wessels with $M = \emptyset$ and $\alpha < 1$ (cf. [46]) but is weaker for discount factors close to 1, since then typically Wessels' stronger version will not hold.

Hordijk shows in [19] that for bounded μ (1.1) is equivalent to the so-called "simultaneous Doeblin condition". This implies [19], that the MCs induced by the stationary (multi-)policies are strongly ergodic, uniformly in the stationary policies. This condition is well-known to fail for most queueing applications. However, for general $\mu \geq 1$ a similar result has been shown to hold in [23, 12]. Under additional continuity assumptions, (1.1) implies the existence of constants $\beta < 1$, c, such that

$$\sup_{i \in \mathsf{E}} \mu_i^{-1} \sum_{j \in \mathsf{E}} \left| P_{ij}^n(\pi) - \Pi_{ij}(\pi) \right| \mu_j \leq c\beta^n, \qquad \forall\, \pi, \tag{1.2}$$

where $\Pi(\pi)$ denotes the stationary probability matrix corresponding to π. This allows a simple derivation of all existence results for the average optimality criterion.

In fact due to this implication, the proof techniques in [15] directly extend to the problem considered in this paper, except for the convergence of the value iteration algorithm, which is more involved when allowing for general μ vectors.

As far as we know, there are only two other recent papers on average optimality in countable state Markov games with unbounded rewards. Most closely related to the present paper is the paper [7] by Borkar and Ghosh. The Lyapunov stability condition used there, requires unbounded μ-vectors and is essentially equivalent to (1.1) by results from [42]. However, in [7] a single communicating class assumption is used and the transition kernels are required to have finite supports, but their results are stronger in that the average equilibrium is shown not only to be optimal in expectation, but also with probability 1. On the other hand, [7] does not consider convergence of value iteration nor applications.

Sennott [39] directly analyses the optimality equations and her conditions are formulated likewise. It is not transparent how these conditions will translate to conditions on the transition and reward structure. In the examples she also needs ergodicity under all stationary policies, and the main advantage of her approach seems to be that she does not require that the Laplace-Stieltjes transforms of the transition kernels have to exist in a neighborhood of 0, which we do need for (1.1) to hold. On the other hand, she requires the costs to be bounded from below, which we do not.

We end this introduction by a brief survey of existing literature on stochastic games with bounded cost. The survey by Parthasarathy and Stern in 1977 [31] already contained more than 150 references, so we will not attempt to present the complete state of the art, but will rather mention some of the important steps and a few directions of research in that field which are related to our paper. For a comprehensive and detailed survey on computation algorithms for stochastic games, see Ragahavan and Filar [33].

The foundations of the theory of stochastic zero-sum games were introduced by Shapley in his pioneering 1953 paper [36], that dealt with the total cost criterion (of which the discounted cost is a special case). Only in the 80s the existence of the value for zero-sum games with expected average cost was established, see Mertens and Neyman [27]. They use the asymptotic theory for stochastic games developed by Bewley and Kohlberg in [6], and relate the value of the expected average cost to the limit of the values of discounted games. An alternative approach for proving the existence of the value is given by Maitra and Sudderth [25].

Stochastic games with infinite state space have been studied by Maitra and Sudderth in [25], Majumdar and Sundaram [26], Mertens and Neyman [27], Mertens and Parthasarathy [28], Parthasarathy and Sinha [32], Rieder [34] and many others. Unlike the (infinite) countable case, the question of existence of stationary equilibrium policies for non-countable state spaces is still open, except for some special cases (e.g. [32]), and fairly general conditions are known only for the existence of ϵ-optimal equilibrium policies. The existence of non-stationary sub-game perfect equilibrium for discounted stochastic games with non-countable state space was established in [28]; the existence of stationary correlated equilibria was established for both the discounted [30] and expected average cost [29] having a special ergodic structure. The existence of a value for zero-sum games with infinite state space and average cost was established using transfinite iteration by Maitra and Sudderth in e.g. [25] (general state space).

Finally, we cite some existing references on the value iteration algorithm applied to the expected average cost: [17,45].

2. α-DISCOUNT AND AVERAGE OPTIMALITY FOR N-PERSON NON-COOPERATIVE GAMES.

We will first describe the N-person non-cooperative Markov game to be studied here. The game is observed at discrete time points to be in some state, say i, from a countable set of states E. $A^n(i)$ is a compact metric set denoting the actions available to player n, $n \in \{1, \ldots, N\}$ and

$$A(i) \triangleq \times_{n=1}^N A^n(i).$$

Let $\mathcal{S}^n(i)$ denote the space of probability measures on $\mathcal{B}(A^n(i))$, the Borel subsets of $A^n(i)$, endowed with the weak topology; and

$$\mathcal{S}(i) \triangleq \times_{n=1}^N \mathcal{S}^n(i).$$

It then follows, that $\mathcal{S}^n(i)$ is a linear Hausdorff compact convex metric space and hence $\mathcal{S}(i)$. And so

$$\mathcal{S} \triangleq \times_{i\in\mathsf{E}} \mathcal{S}(i)$$

is a metrisable, convex, compact set in the weak topology of componentwise convergence (cf. [15]).

If the system is in state i at time t, player n selects a decision rule $\pi^n(i) \in \mathcal{S}^n(i)$ independently of the other players, but this selection may depend on the history $\mathcal{H}(t) = \{i_0, a_0, \ldots, i_{t-1}, a_{t-1}, i_t\}$ of previous states and action vectors as well as on the present state, and so this is a game of perfect recall. If the outcome of all players' decisions is the vector $a \in A(i)$, then player n earns a reward $r_{i,a}^n$ and the system moves to state j with probability P_{iaj}.

The immediate rewards are represented by the vector

$$r_{ia} = \{r_{ia}^1, \ldots, r_{ia}^N\}.$$

As a shorthand notation we write

$$r_i^n(\pi) = \int_{A(i)} r_{ia}^n d\pi(a), \quad \pi \in \mathcal{S}(i),$$

for the expected immediate reward under decision rule π; $P_{ij}(\pi)$ is defined similarly.

We will use the following standard assumption.

ASSUMPTION 1: *P_{iaj} and r_{ia} are continuous on $A(i)$.*

This implies that $r_i^n(\pi)$ and $P_{ij}(\pi)$ are continuous on $\mathcal{S}(i)$ (cf. [15]).

Next we will indicate the various multi-policies to be considered. A multi-policy R is a sequence of decision rules $\{\pi_0, \ldots, \pi_t, \ldots\}$, where the decision rule π_t at time t may depend on $\mathcal{H}(t)$ (see [13] for a precise construction). We will omit the suffix "multi", whenever it is clear that we denote a simultaneous policy for n players. For a semi-Markov policy π_t only depends on the initial state and the state at time t; it is Markov, if in addition there is no dependence on the initial state. A policy is stationary if $\pi_t = \pi_0, \ \forall\, t$, and it is deterministic if π_t is a degenerate distribution.

Denote the set of all policies, all semi-Markov policies and all Markov policies by $\mathcal{C}$, $\mathcal{C}_{sM}$ and $\mathcal{C}_M$ respectively. Clearly the stationary policies can be identified with the stationary decision rules and hence by abuse of notation π and $\mathcal{S}$ will also denote a stationary policy and set of stationary policies respectively. Finally $f \in \mathcal{S}$ and $\mathcal{F} \subset \mathcal{S}$ will denote a stationary deterministic policy and the set of such policies. A superscript n will be used to denote the policy or decision rule used, by player n. Define the total expected α-discounted reward $V_i^{\alpha,n}(R)$ of player n when the initial state is i and multi-policy $R \in \mathcal{C}$ is used by

$$V_i^{\alpha,n}(R) \triangleq \mathbb{E}_R\Big(\sum_{t\geq 0} \alpha^t r^n_{X(t),Y(t)} \big| X(0)=i\Big),$$

where $\{X(t),Y(t)\}_{t=0,1,\ldots}$ is the stochastic process of realised states and actions and $\mathbb{E}_R$ the expectation operator given the players' policy R. The long-run average reward $g_i^n(R)$ per unit-time and the strong long-run average reward $g_i^{s,n}(R)$ are given by:

$$g_i^n(R) \triangleq \liminf_{t\to\infty} \frac{1}{t+1}\mathbb{E}_R\Big(\sum_{l=0}^{t} r^n_{X(l),Y(l)} \big| X(0)=i\Big)$$
$$g_i^{s,n}(R) \triangleq \limsup_{t\to\infty} \frac{1}{t+1}\mathbb{E}_R\Big(\sum_{l=0}^{t} r^n_{X(l),Y(l)} \big| X(0)=i\Big).$$

Let us write

$$\mathcal{C}^{-n}(R) = \{\tilde{R} \in \mathcal{C} : \tilde{R}^k = R^k, k \neq n\}$$

as the set of policies, such that the restriction to the k^{th} player's policy coincides with R^k, $k \neq n$. Then R^* is said to be an *α-discounted equilibrium point of policies* (α-DEP), if for each player n

$$V_i^{\alpha,n}(R^*) \geq V_i^{\alpha,n}(R), \quad \forall i \in \mathsf{E}, \forall R \in \mathcal{C}^{-n}(R^*).$$

R^* is an *average equilibrium point of policies* (AEP), if for each player n

$$g_i^n(R^*) \geq g_i^n(R), \quad \forall i \in \mathsf{E}, \forall R \in \mathcal{C}^{-n}(R^*).$$

The definition of a strong average equilibrium point (s-AEP) is completely analogous. We will need that AEPs are also s-AEPs and vice versa, for analysing the 2-person zero-sum game as a special case of the 2-person non-cooperative game.

As has been said in the Introduction our analysis is based on using normed linear operators. The space of interest will be V_μ, the Banach space of vectors in E with a finite μ-norm, for a given positive real-valued vector $\mu \geq 1$ on E, i.e.

$$\|x\|_\mu \triangleq \sup_{i\in\mathsf{E}} \frac{|x_i|}{\mu_i}.$$

If $A : \mathsf{V}_\mu \to \mathsf{V}_\mu$ is a bounded linear operator, then the associated operator norm is equal to

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

We will need a slightly stronger assumption than assumption 1 to achieve the desired continuity, when working in this Banach space.

Assumption 1^*:

i) r_{ia} *is continuous on* $A(i)$ *and* μ*-bounded, i.e.*

$$\sup_{i \in \mathsf{E}} \sup_{a \in A(i)} \frac{|r_{ia}|}{\mu_i} < \infty.$$

ii) μ*-continuity of the transition probabilities, i.e. if* $a(n) \to a$ *in* $A(i)$ *then*

$$\lim_{n \to \infty} \sum_{j \in \mathsf{E}} |P_{ia(n)j} - P_{iaj}| \mu_j \to 0.$$

Note, that assumption 1^* implies assumption 1 and continuity of $P(\pi)$ on $\mathcal{S}$. Furthermore, μ-continuity is weaker than continuity in operator norm. The latter would not be satisfied in most queueing applications, since it implies $\sup_i |P_{ia(n)i} - P_{iai}| \to 0$ and is hence undesirably strong.

We will next state our basic assumption, which is (1.1) from the Introduction.

Assumption $2(\alpha)$: *there is a finite set* $M \subset \mathsf{E}$ *and a constant* $\beta < 1$*, such that*

$$\alpha \sum_j {}_M P_{iaj} \mu_j \leq \beta \mu_i, \quad \forall a \in A(i), \forall i \in \mathsf{E},$$

where

$${}_M P_{iaj} = \begin{cases} P_{iaj}, & j \notin M \\ 0, & j \in M. \end{cases}$$

For $\alpha = 1$ this condition is referred to as *μ-uniform geometric recurrence* (cf. [11, 12]). It implies (1.1) evidently. It can easily be shown, that 2(1) never holds for $M = \emptyset$, if there exists some stationary multi-policy such that the corresponding Markov chain is recurrent.

2.1 Existence of stationary α-DEPs for the infinite horizon problem

We will first analyse the α-discounted reward case and so we fix $\alpha \in (0,1)$. The following is an easy consequence of our conditions.

Lemma 2.1. *Under assumptions* 1^* *and* $2(\alpha)$ *it holds that* $V_i^{\alpha,n}(\pi)$ *is continuous on* $\mathcal{S}$*,* $i \in \mathsf{E}$*, and* $\sup_{\pi \in \mathcal{S}} \|V^{\alpha,n}(\pi)\|_\mu < \infty$ *for player* n*.*

Proof First note that assumption $2(\alpha)$ implies that

$$\sup_{\pi \in \mathcal{S}} \|P(\pi)\|_\mu < \infty. \tag{2.1}$$

Consequently μ-continuity of $P(\pi)$ implies inductively μ-continuity of the t^{th} iterate $P^t(\pi)$ of $P(\pi)$ by dominated convergence.

Denote by $\tau_M \geq 1$ the first hitting time of set M and write ${}_M P^t(\pi) = \left({}_M P(\pi)\right)^t$, so that ${}_M P^0(\pi) = I$ for I the identity matrix. By first-entrance last-exit decomposition on set M (cf. [9], p.

47) we find

$$\alpha^t \sum_j P^t_{ij}(\pi)\mu_j = \alpha^t \sum_j {}_M P^t_{ij}(\pi)\mu_j + \sum_{m,m'\in M} \sum_{k=1}^t \mathbb{P}_\pi\{\tau_M = k, X(\tau_M) = m\}\times$$
$$\times \sum_{u=0}^{t-k} \alpha^{t-u} P^{t-k-u}_{mm'}(\pi)\alpha^u \sum_j {}_M P^u_{m'j}\mu_j$$
$$\leq \beta^t \mu_i + \max_{m'\in M} \mu_{m'} \sum_{k=1}^t \mathbb{P}_\pi\{\tau_M = k\} \sum_{u=0}^{t-k} \alpha^{t-u}\beta^u$$
$$\leq \big\{\beta^t + \max_{m'\in M} \mu_{m'} \sum_{u=0}^{t-1} \alpha^{t-u}\beta^u\big\}\mu_i. \tag{2.2}$$

Hence $\sup_{\pi\in\mathcal{S}} \|\sum_{t\geq 0} \alpha^t P^t(\pi)\|_\mu < \infty$. (2.2) implies that $\sum_{t=0}^T \alpha^t P^t_{ij}(\pi)\mu_j$ converges to $\sum_{t\geq 0} \alpha^t P^t_{ij}(\pi)\mu_j$ uniformly in $\pi \in \mathcal{S}$, as $T \to \infty$. As a consequence $\sum_{t\geq 0} \alpha^t \sum_j P^t_{ij}(\pi)\mu_j$ is continuous in π.

Because of the continuity and μ-boundedness of the rewards, $V^{\alpha,n}(\pi)$ is μ-bounded uniformly in π and it is elementwise continuous in π. To achieve this we use the (generalised) dominating convergence theorem (cf. [35], Proposition 11.18) and Fubini-Tonelli's theorem. **QED**

As in [15] we will obtain the existence of a stationary α-DEP through the optimality equation, but this requires the following fixed point result. (π^{-n}, η) stands for the stationary policy with π^k, $k \neq n$, the policy of player k, and η the policy of player n.

LEMMA 2.2 *Let $v^n(\pi) \in \mathsf{V}_\mu$, $\pi \in \mathcal{S}$, be continuous in π for each player n. Define*

$$w^n_i(\pi^{-n}, \eta) \triangleq r^n_i(\pi^{-n}, \eta) + \alpha \sum_j P_{ij}(\pi^{-n}, \eta) v^n_j(\pi).$$

Then there exists $\tilde\pi \in \mathcal{S}$, such that for each player n and all $i \in \mathsf{E}$

$$w^n_i(\tilde\pi^{-n}, \tilde\pi^n) = \max_{\eta\in\mathcal{S}^n(i)} \big\{r^n_i(\tilde\pi^{-n}, \eta) + \alpha \sum_j P_{ij}(\tilde\pi^{-n}, \eta) v^n_j(\tilde\pi)\big\}. \tag{2.3}$$

Proof Denote by $2^{\mathcal{S}}$ the class of all non-empty subsets of $\mathcal{S}$ and let

$$\Phi^n(\pi) \triangleq \big\{\eta \in \mathcal{S}^n : w^n_i(\pi^{-n}, \eta) = \max_{\theta\in\mathcal{S}^n(i)} w^n_i(\pi^{-n}, \theta), \quad \forall i \in \mathsf{E}\big\}.$$

$\Phi^n(\pi)$ is not empty, because of the μ-continuity of $P(\pi)$, the compactness of $\mathcal{S}^n(i)$ and the conditions imposed on $v^n(\pi)$. Define the mapping $\Phi : \mathcal{S} \to 2^{\mathcal{S}}$ by

$$\Phi(\pi) = \times_{n=1}^N \Phi^n(\pi).$$

As in [15] Φ is an upper-semicontinuous point-to-convex-set mapping of the convex compact set $\mathcal{S}$ into itself. From what has been said before, the conditions of Glickberg's [18] extension of the Kakutani fixed point theorem are satisfied and so there exists a fixed point $\tilde\pi \in \Phi(\tilde\pi)$. **QED**

The result on the existence of stationary α-DEPs easily follows from the existence of stationary optimal policies in Markov decision chains satisfying assumptions 1^* and $2(\alpha)$ (cf. [11]) by combination with lemma 2.2. [11] then implies

THEOREM 2.3 *Under assumptions* 1^* *and* $2(\alpha)$ *the following hold:*

i) $\hat{\pi}$ *is a stationary* α-DEP *if and only if* $V^\alpha(\hat{\pi})$ *satisfies the optimality equation: for all* n

$$V_i^{\alpha,n}(\hat{\pi}) = \max_{\eta \in \mathcal{S}^n(i)} \Big\{ r_i^n(\hat{\pi}^{-n}, \eta) + \alpha \sum_j P_{ij}(\hat{\pi}^{-n}, \eta) V_i^{\alpha,n}(\hat{\pi}) \Big\}, \quad \forall i \in \mathsf{E}; \tag{2.4}$$

ii) there is a stationary α-DEP.

Proof First note that similarly to (2.2), one can show the existence of constants $\gamma < 1$ and $c < \infty$, such that for all t

$$\sup_{\pi_0,\ldots,\pi_t} \alpha^t \|P(\pi_0)\cdots P(\pi_t)\|_\mu \le c\gamma^t. \tag{2.5}$$

Assume that $V^\alpha(\hat{\pi})$ satisfies (2.4). Then for $\eta = \hat{\pi}^n$ there is equality in equation (2.4). When all other players use policy $\hat{\pi}^{-n}$, player n is faced with an MDP and may thus restrict to Markov policies (see e.g. [19]). Let $(\pi_0^n, \pi_1^n, \ldots)$ be any Markov policy for player n. By iteration of (2.4), we find for any T that

$$\sum_{t=0}^{T} \alpha^t P(\hat{\pi}^{-n}, \pi_0^n)\cdots P(\hat{\pi}^{-n}, \pi_{t-1}^n) r(\hat{\pi}^{-n}, \pi_t^n) + \alpha^{T+1} P(\hat{\pi}^{-n}, \pi_0^n)\cdots P(\hat{\pi}^{-n}, \pi_T^n) V^{\alpha,n}(\hat{\pi}) \le V^{\alpha,n}(\hat{\pi}).$$

From the fact that $\|V^{\alpha,n}(\hat{\pi})\|_\mu < \infty$ and that

$$\alpha^{T+1} P(\hat{\pi}^{-n}, \pi_0^n)\cdots P(\hat{\pi}^{-n}, \pi_T^n) V^{\alpha,n}(\hat{\pi}) \to 0, \quad T \to \infty,$$

by virtue of (2.5), it follows that

$$V^{\alpha,n}\Big((\hat{\pi}^{-n}, \pi_0^n), \ldots, (\hat{\pi}^{-n}, \pi_t^n), \ldots\Big) \le V^{\alpha,n}(\hat{\pi}),$$

for any n and any Markov policy $(\pi_0^n, \pi_1^n, \ldots)$ for player n. As a consequence $\hat{\pi}$ is a stationary α-DEP.

Next assume that $\hat{\pi}$ is a stationary α-DEP and that $V^{\alpha,n}(\hat{\pi})$ does not satisfy the optimality equation (2.4) for some player n. Then there is a state, i say, such that

$$V_i^{\alpha,n}(\hat{\pi}) < \max_{\eta \in \mathcal{S}^n(i)} \Big\{ r_i^n(\hat{\pi}^{-n}, \eta) + \alpha \sum_j P_{ij}(\hat{\pi}^{-n}, \eta) V_j^{\alpha,n}(\hat{\pi}) \Big\}. \tag{2.6}$$

Let $\hat{\eta}$ be the action for player n for which the maximum in the right-handside of the above inequality is attained. Then again by iteration of (2.6) and using the bounds (2.5) it follows that

$$V_i^{\alpha,n}(\hat{\hat{\pi}}) > V_i^{\alpha,n}(\hat{\pi}),$$

for $\hat{\hat{\pi}}^{-n} = \hat{\pi}^{-n}$ and

$$\hat{\hat{\pi}}^n = \begin{cases} \hat{\eta}, & j = i \\ \hat{\pi}_j^n, & j \ne i. \end{cases}$$

This proves i). For the proof of ii) take $v(\pi) = V^\alpha(\pi)$ in the statement of lemma 2.2. **QED**

2.2 Existence of Markov α-DEPs for the finite horizon problem

This subsection briefly discusses the finite horizon problem as a special case of the infinite horizon case, by an appropriate change of the state space so as to include time.

Consider the discounted cost for a finite horizon T, with some finite nonnegative discount factor α $(0 \le \alpha < \infty)$, i.e., where in the definition of $V_i^{\alpha,n}(R)$ we sum t only between 0 and $T < \infty$. Assume that assumption 1^* holds, and that assumption $2(\overline{\alpha})$ holds for some $0 < \overline{\alpha} < 1$ (not necessarily the same as α). Then there exist a Markov α-DEP. To show this, we consider a new infinite horizon Markov game with state space $\overline{\mathsf{E}} = \mathsf{E}^{T+1} \cup \{\gamma\}$, where γ is an absorbing state (see below). The action spaces are unchanged, the rewards are given by

$$\overline{r}_{(i,t),a} = (\alpha/\overline{\alpha})^t r_{i,t}, \quad t = 0, ..., T, \qquad \overline{r}_{\gamma,a} = 0.$$

The transition probabilities are

$$\overline{P}_{\gamma,a,\gamma} = 1, \quad \overline{P}_{(i,0),a,\gamma} = 1, \quad \overline{P}_{(i,t),a,(j,s)} = P_{iaj} \quad \text{for } s = t-1,\ t \ge 0;$$

and all other transitions have zero probability. The discount factor is $\overline{\alpha}$. It is easily seen that the new model is equivalent to the original one; there is a natural one-to-one correspondence between the policies in the two models, such that corresponding policies have the same costs. In particular, Markov policies in the finite horizon model are equivalent to stationary policies in the new infinite horizon problem in an obvious way. Finally, we note that assumptions 1^* and $2(\overline{\alpha})$ hold in the new model (the latter follows by considering $\overline{M} = \{(i,t) : i \in M\}$, $\overline{\mu}_{(i,t)} = \mu_i$, and $\mu(\gamma) = 1$). Therefore, we may conclude that the existence of stationary α-DEPs for the infinite horizon problem, established in the previous subsection, translates into the existence of Markov α-DEPs for our finite horizon problem.

2.3 Existence of stationary AEPs

This subsection treats the average reward case as a limiting case of the α-discounted reward model. Although we will not explicitly impose any conditions on the number of closed classes under the stationary policies, we do require that it be continuous. This essentially implies that the number of positive recurrent classes is a constant function of the stationary policies. This is clearly a major restriction compared to the case of only 1 player, where it is sufficient to require this condition for the class $\mathcal{F}$ of stationary, deterministic policies only. This roughly implies, that the problem consists of a number of irreducible subproblems, which can be solved separately. The outcome however will not make precise how to reach the best possible irreducible set of states from transient states. This is our motivation for preferring to study the general multichain model. For a detailed discussion of the problems arising when this continuity condition is lacking, we refer to [15, 16].

Essential for our discussion are the following results from [12]. $\nu(\pi)$ will denote the number of closed classes in the Markov chain resulting from policy π. Further

$$\Pi_{ij}(\pi) = \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P_{ij}^t(\pi),$$

is the corresponding stationary matrix. Let $\alpha = 1$.

THEOREM 2.4 *Suppose that assumptions* 1^* *and* $2(1)$ *hold. Let* $\mathcal{C}^*$ *stand for either* $\mathcal{S}$ *or* $\mathcal{F}$. *If* ν *is a continuous function on* $\mathcal{C}^*$, *then there are constants* $\gamma < 1$ *and* c, *such that for all* $\pi \in \mathcal{C}^*$

$$\|P^t(\pi) - \Pi(\pi)\|_\mu \le c \cdot \gamma^t,$$

provided the Markov chain resulting from π is aperiodic.

By using a transformation, this theorem can be used to show [10,41] that the α-discounted value function $V^{\alpha,n}(\pi)$ admits a Laurent expansion for all $\pi \in \mathcal{C}^*$ in some uniform disc around $\alpha = 1$.

THEOREM 2.5 *Suppose, that assumptions 1^* and $2(1)$ hold. If ν is continuous on $\mathcal{C}^*$, then there is an $\epsilon > 0$, such that for all $\pi \in \mathcal{C}^*$*

$$V^{\alpha,n}(\pi) = \frac{1}{1-\alpha}\Pi(\pi)r^n(\pi) + \frac{1}{\alpha}\sum_{k\geq 0}\left(1-\frac{1}{\alpha}\right)^k D^{k+1}(\pi)r^n(\pi), \quad |1-\alpha^{-1}| < \epsilon, \tag{2.7}$$

where $D^{k+1}(\pi)$ is the k^{th} iterate of the deviation matrix

$$D(\pi) = \lim_{\alpha\uparrow 1}\sum_{t\geq 0}\left(P^t(\pi) - \Pi(\pi)\right)\alpha^t.$$

Moreover $\sup_{\pi\in\mathcal{C}^*}\|D(\pi)\|_\mu < \infty$ and $D(\pi)$ is μ-continuous on $\mathcal{C}^*$.

The following average optimality result can be derived.

THEOREM 2.6 *Suppose that assumptions 1^* and $2(1)$ hold and that ν is continuous on $\mathcal{S}$. Then:*
i) π is an AEP *if and only if (g,v) given by*

$$g^n = \Pi(\pi)r^n(\pi) \tag{2.8a}$$
$$v^n = D(\pi)r^n(\pi) \tag{2.8b}$$

solves the following pair of optimality equations for each player n:

$$g_i^n = \max_{\eta\in\mathcal{S}^n(i)}\sum_j P_{ij}(\pi^{-n},\eta)g_j^n, \qquad i \in \mathsf{E} \tag{2.9a}$$

$$v_i^n = \max_{\eta\in\mathcal{S}^{n,*}(i)}\left\{r_i^n(\pi^{-n},\eta) - g_i^n + \sum_j P_{ij}(\pi^{-n},\eta)v_j^n\right\}, \quad i \in \mathsf{E} \tag{2.9b}$$

with

$$\mathcal{S}^{n,*}(i) = \left\{\eta \in \mathcal{S}^n \;\middle|\; \begin{array}{l} g^n = P(\pi^{-n},\eta)g^n \\ i \text{ is positive recurrent under } (\pi^{-n},\eta)\end{array}\right\}. \tag{2.9c}$$

*ii) Any limit-point π_∞ of a sequence of α-*DEPs *$\{\pi_\alpha\}$, for $\alpha \to 1$, is an* AEP. *It is a solution to (2.9a,b) with (g,v) given by (2.8) for $\pi = \pi_\infty$ and*

$$\mathcal{S}^{n,*}(i) = \left\{\eta \in \mathcal{S}^n(i) \,|\, g_i^n = \sum_j P_{ij}(\pi^{-n},\eta)g_j^n\right\}. \tag{2.9d}$$

The equivalence result in i) is not given in explicit form even for the Markov decision model, and so we will give extensive details here. Note that a limit point of α-DEPs satisfies the *stronger* optimality equations (2.9a,b,d), which are the ones that are usually considered.

Proof We prove i). First assume that π is an AEP. Fix player n. For simplicity we will suppress the dependence on π^{-n} in our notation.

For any policy η of player n, the Laurent expansion of $V^{\alpha,n}(\eta)$ exists by theorem 2.5 and it has the form

$$V^{\alpha,n}(\eta) = (1+\rho)\{\frac{g^n(\eta)}{\rho} + \sum_{k\geq 0}(-\rho)^k x^{n,k}(\eta)\},$$

where $\rho = (1-\alpha)/\alpha$ is the interest rate and $x^{n,k}(\eta), g^n(\eta) \in \mathsf{V}_\mu$. Note that $v^n = x^{n,0}(\pi^n)$.

Theorem 2.5 also implies that the inverse of $\big(I - (1+\rho)^{-1}P(\eta)\big)$ exists as an operator on V_μ for small values of ρ and it is equal to

$$\big(I - \frac{1}{1+\rho}P(\eta)\big)^{-1} = \frac{1+\rho}{\rho}\big\{\Pi(\eta) + \sum_{k\geq 1}(-1)^{k-1}\rho^k D^k(\eta)\big\}.$$

Write $x^n = V^{\alpha,n}(\pi^n)$. For any stationary policy η of player n

$$\lim_{\rho\downarrow 0}\rho\big(V^{\alpha,n}(\eta) - x^n\big) = g^n(\eta) - g^n(\pi^n). \tag{2.10}$$

For ρ fixed, we have

$$\begin{aligned}
\rho\big(V^{\alpha,n}(\eta) - x^n\big) &= (1+\rho)\Big\{\Pi(\eta) + \sum_{k\geq 1}(-1)^{k-1}\rho^k D^k(\eta)\Big\}\Big\{r^n(\eta) + \frac{1}{1+\rho}P(\eta)x^n - x^n\Big\}\\
&= (1+\rho)\Big\{\Pi(\eta) + \rho D(\eta)\Big\}\Big\{\frac{1}{\rho}\big[P(\eta)g^n(\pi^n) - g^n(\pi^n)\big] + r^n(\eta) + P(\eta)x^{n,0}(\pi^n)\\
&\qquad - x^{n,0}(\pi^n) - g^n(\pi^n)\Big\} + \ \rho\text{-terms of order} \geq 1\\
&= \Pi(\eta)\Big\{r^n(\eta) + P(\eta)x^{n,0}(\pi^n) - x^{n,0}(\pi^n) - g^n(\pi^n)\Big\}\\
&\qquad - D(\eta)\Big\{g^n(\pi^n) - P(\eta)g^n(\pi^n)\Big\} + \ \rho\text{-terms of order} \geq 1.
\end{aligned} \tag{2.11}$$

Suppose that (2.9a) is not satisfied. Then there is an η, a state i' and $\epsilon > 0$ for which

$$g^n_{i'}(\pi^n) \leq \sum_j P_{i'j}(\eta)g^n_j(\pi^n) - \epsilon,$$

and for which there is equality in (2.9a) for η in all other states (that is, there is a one state improvement). If i' is positive recurrent under η, then

$$\big[\Pi(\eta)\big\{P(\eta)g^n(\pi^n) - g^n(\pi^n)\big\}\big]_{i'} \geq \Pi_{i'i'}(\eta)\epsilon,$$

which contradicts the fact that the lefthandside of this expression is 0. Suppose that i' is transient. Then $\Pi(\eta)\big\{r^n(\eta) + P(\eta)x^{n,0}(\pi^n) - x^{n,0}(\pi^n) - g^n(\pi^n)\big\} = 0$, and so (2.10) and (2.11) imply

$$\begin{aligned}
g^n_{i'}(\eta) - g^n_{i'}(\pi^n) &= -\big[D(\eta)\big(g^n(\pi^n) - P(\eta)g^n(\pi^n)\big)\big]_{i'}\\
&\geq D_{i'i'}(\eta)\epsilon \geq \epsilon > 0,
\end{aligned}$$

since $D_{i'i'}(\eta) = \sum_{t\geq 0}P^t_{i'i'}(\eta) \geq 1$. This contradicts the optimality of π^n. Finally we assume, that (2.9a) is satisfied, but (2.9b,c) not. Then there exists η, a state i' that is positive recurrent under η and $\epsilon > 0$, for which

$$r^n_{i'}(\eta) + \big[P(\eta)x^{n,0}(\pi^n)\big]_{i'} - x^{n,0}(\pi^n)_{i'} - g^n_{i'}(\pi^n) \geq \epsilon$$

and there is equality in (2.9b) for all other positive recurrent states under η. (2.10) and (2.11) imply

$$g^n_{i'}(\eta) - g^n_{i'}(\pi^n) = \big[\Pi(\eta)\{r^n(\eta) + P(\eta)x^{n,0}(\pi^n) - x^{n,0}(\pi^n) - g^n(\pi^n)\}\big]_{i'} \geq \Pi_{i'i'}(\eta)\epsilon > 0.$$

Consequently (g, v) given in (2.8) solve (2.9a,b,c).

For the reverse implication let $g^n_{i'}(\eta) > g^n_{i'}(\pi^n) + \epsilon$ for some state i' and $\epsilon > 0$. Analogously to our previous reasoning there must be equality in (2.9a) for all states that are positive recurrent under η. On these positive recurrent states i consequently

$$v^n_i \geq r^n(\eta) - g^n_i + \sum_j P_{ij}(\eta) v^n_j.$$

Similarly to the proof of theorem 2.8 it can be shown that then necessarily $g^n_i(\eta) \leq g^n_i(\pi^n)$ for i positive recurrent under η. Let i' be transient under η. Then (2.9a) must be satisfied with equality by η for all transient states j with $D_{i'j}(\eta) > 0$. But $D_{i'j}(\eta) > 0$ if and only if j can be reached from i' with positive probability under η.

Let C_1, ...,C_ν be the positive recurrent classes under η and T the transient states. From the foregoing

$$\begin{aligned} g^n_{i'}(\pi^n) &= \sum_l \sum_{j \in C_l} P_{i'j}(\eta) g^n_j(\pi^n) + \sum_{j \in T} P_{i'j}(\eta) g^n_j(\pi^n) \\ &= \sum_l \sum_{j \in C_l} \sum_{t \geq 1} \left[{}_T P^{t-1}(\eta) P(\eta)\right]_{i'j} g^n_j(\pi^n) \\ &\geq \sum_l \sum_{j \in C_l} \sum_{t \geq 1} \left[{}_T P^{t-1}(\eta) P(\eta)\right]_{i'j} g^n_j(\eta) = g^n_{i'}(\eta), \end{aligned}$$

a contradiction. This proves i).

For ii) note that assumption 2(1) implies assumption 2(α) in a neighborhood of 1, so that the existence of stationary α-DEPs is guaranteed. The corresponding α-discounted value functions satisfy the optimality equation (2.4) and have a Laurent expansion given by (2.7), for all $\alpha \geq \hat{\alpha}$. Since $\mathcal{S}$ is a compact set, $\{\pi_\alpha\}$ has a convergent subsequence for $\alpha \uparrow 1$ and hence there is at least one limit-point $\pi_\infty \in \mathcal{S}$. Say that $\pi_{\alpha(k)} \to \pi_\infty$, $k \to \infty$. For $\alpha_k \geq \hat{\alpha}$ we find from (2.7)

$$(1 - \alpha(k)) V^{\alpha,n}(\pi^{-n}_{\alpha(k)}, \eta) = g^n(\pi^{-n}_{\alpha(k)}, \eta) + \frac{1 - \alpha(k)}{\alpha(k)} x^{0,n}(\pi^{-n}_{\alpha(k)}, \eta) + o(1 - \alpha(k)), \qquad (2.12)$$

where $g^n(\pi^{-n}_{\alpha(k)}, \eta) = \Pi(\pi^{-n}_{\alpha(k)}, \eta) r^n(\pi^{-n}_{\alpha(k)}, \eta)$ is the average expected reward under $(\pi^{-n}_{\alpha(k)}, \eta)$. Using that

$$V^{\alpha,n}(\pi_{\alpha(k)}) \geq V^{\alpha,n}(\pi^{-n}_{\alpha(k)}, \eta),$$

plugging in the expressions (2.12), and taking the limit $k \to \infty$ it follows that π_∞ is an AEP. The fact that π^∞ solves (2.9a,b,d) with (g, v) of the form (2.8) for $\pi = \pi_\infty$, follows by equating terms of the Laurent expansion of $\rho\big(V^{\alpha,n}(\pi^{-n}_\infty, \eta) - V^{\alpha,n}(\pi_\infty)\big)$ in a similar way as has been done in the proof of i).

QED

In case the Markov chains resulting from the stationary policies are all unichain, the average reward vector in (2.9a) is a constant vector and so the optimality equations reduces to equations (2.9b,c) and (2.9b,d) respectively.

We will next consider the *strong* average optimality case: under our conditions

$$g^n(\pi) = \liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t(\pi) r^n(\pi) = \limsup_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t(\pi) r^n(\pi), \quad \pi \in \mathcal{S}, \tag{2.13}$$

by virtue of theorem 2.5, so that strong average optimality and average optimality are equivalent when restricting to the class $\mathcal{S}$. It will not be surprising that this equivalence extends to class $\mathcal{C}$ of all policies.

A result of this kind is necessary for going from the N-person non-cooperative game to the two-person zero sum game. For the proof of our Theorems for the two-person zero sum game we need a maximisation in (2.9b) of the *strong* optimality equations over *all* randomised decisions of player n.

Introduce the following system of vectors $\tilde{g}, \tilde{v} \in \mathsf{V}_\mu^N$: for some $\pi \in \mathcal{S}$ and any player n it holds that

$$\tilde{g}_i^n = \max_{\eta \in \mathcal{S}^n(i)} \sum_j P_{ij}(\pi^{-n}, \eta) \tilde{g}_j^n \qquad i \in \mathsf{E} \tag{2.14a}$$

$$\tilde{v}_i^n = \max_{\eta \in \mathcal{S}^n(i)} \left\{ r_i^n(\pi^{-n}, \eta) - g^n + \sum_j P_{ij}(\pi^{-n}, \eta) v_j^n \right\}, \quad i \in \mathsf{E}. \tag{2.14b}$$

We will show the existence of a solution to system (2.14): it uses an elegant argument due to Schäl [37]. Although this follows directly from [41, Ch. 7], we will give the arguments here for completeness of the exposition, because they have not been published.

LEMMA 2.7 *Let $(g, v) \in \mathsf{V}_\mu^{2N}$ be a solution of (2.9a,b,d) for some policy $\pi \in \mathcal{S}$. Then under assumptions 1^* and 2(1) there exists $\hat{v} \in \mathsf{V}_\mu^N$, such that $(g, \hat{v})$ is a solution of (2.14) for the same policy π.*

Proof Consider a transformed Markov game with immediate rewards $\hat{r}_{ia} = r_{ia} - g$. We denote by $\hat{g}(\pi)$ the expected average reward under policy π in this game.

Fix the policy of player k, $k \neq n$, and set it equal to π^k. We solve the Markov decision problem for player n in the transformed game. As in theorem 2.6 there is a pair $(\hat{g}^n, \hat{v}^n) \in \mathsf{V}_\mu^2$ solving (2.9a,b,d) for player n and some policy $\hat{\eta} \in \mathcal{S}^n$, with $\hat{g}^n = \hat{g}^n(\pi^{-n}, \hat{\eta})$. Clearly

$$g^n \geq g^n(\pi^{-n}, \hat{\eta}) = \Pi(\pi^{-n}, \hat{\eta}) r^n(\pi^{-n}, \hat{\eta}).$$

Multiplication by $\Pi(\pi^{-n}, \hat{\eta})$ yields

$$\Pi(\pi^{-n}, \hat{\eta}) g^n \geq \Pi(\pi^{-n}, \hat{\eta}) r^n(\pi^{-n}, \hat{\eta}),$$

so that

$$\hat{g}^n = \Pi(\pi^{-n}, \hat{\eta}) \big(r^n(\pi^{-n}, \hat{\eta}) - g \big) \leq 0.$$

On the other hand, $\hat{g}^n$ is the optimal reward for player n in the above transformed Markov decision problem, and hence

$$\hat{g}^n \geq \hat{g}^n(\pi^n) = \Pi(\pi^n) \big(r^n(\pi) - g \big) = 0.$$

Combination gives $\hat{g}^n = 0$, so that $\hat{g}^n = P(\pi^{-n}, \eta) \hat{g}^n$ for all $\eta \in \mathcal{S}^n$. It follows that the maximisation in (2.9b,d) for the pair $(g^n, \hat{v}^n)$ is over $\mathcal{S}^n(i)$, thus completing the proof. **QED**

The following result is now easily proved.

THEOREM 2.8 *Suppose that assumptions* 1^* *and* $2(1)$ *hold. Any stationary* AEP *solving (2.9a,b,d) is also an* s-AEP. *In particular any limit-point of* α-DEPs, *for* $\alpha \to 1$, *is an* s-AEP.

Proof Let $\pi \in \mathcal{S}$ be an AEP satisfying (2.9a,b,d) for some $(g, v) \in \mathsf{V}_{\mu}^{N}$. Any limit-point of α-DEPs is of this form. It is sufficient to show that $g \geq g(\pi^{-n}, R^n)$ for $R^n = (\pi_0^n, \pi_1^n, \ldots)$ a Markov policy for player n. This follows from [21]. Note, that this paper considers the case of countable state *and* action spaces, but the results there are easily generalised to the countable state space and compact action space case.

Similar to the proof of lemma 2.1 we can show the existence of a constant c, such that

$$\|P(\pi^{-n}, \pi_0^n) \cdots P(\pi^{-n}, \pi_t^n)\|_{\mu} \leq c, \quad t = 0, 1, \ldots. \tag{2.15}$$

For simplicity we will suppress the dependence on π^{-n} in our notation for the remainder of this proof. By (2.14a) we have $g_i^n \geq \mathbb{E}_{R^n}\big(g_{X(t)}^n | X(0) = i\big)$ for all t, and so iterating (2.14b) we find

$$\begin{aligned}
\mathbb{E}_{R^n}\big(v_{X(t)}^n \mid X(0) = i\big) &\geq \mathbb{E}_{R^n}\big(r_{X(t),Y(t)}^n \mid X(0) = i\big) - \mathbb{E}_{R^n}\big(g_{X(t)}^n \mid X(0) = i\big) \\
&\qquad + \mathbb{E}_{R^n}\big(v_{X(t+1)}^n \mid X(0) = i\big) \\
&\geq \mathbb{E}_{R^n}\big(r_{X(t),Y(t)}^n \mid X(0) = i\big) - g_i^n + \mathbb{E}_{R^n}\big(v_{X(t+1)}^n \mid X(0) = i\big).
\end{aligned}$$

Subtract $\mathbb{E}_{R^n}\big(v_{X(t+1)}^n \mid X(0) = i\big) - g_i^n$, take the summation over $t = 0, \ldots, T$ and divide by $T + 1$, then

$$\frac{1}{T+1} v_i^n - \frac{1}{T+1} \mathbb{E}_{R^n}\big(v_{X(T+1)}^n \mid X(0) = i\big) + g_i^n \geq \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{R^n} r_{X(t),Y(t)}^n.$$

Since $|\mathbb{E}_{R^n}\big(v_{X(T+1)}^n \mid X(0) = i\big)| \leq \tilde{c}\mu_i$ for any T and some constant $\tilde{c}$ by (2.15), we finally infer that

$$\begin{aligned}
g_i^n &= \limsup_{T \to \infty} \left\{ \frac{1}{T+1} v_i^n - \frac{1}{T+1} \mathbb{E}_{R^n}\big(v_{X(T+1)}^n \mid X(0) = i\big) + g_i^n \right\} \\
&\geq \limsup_{T \to \infty} \left\{ \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{R^n}\big(r_{X(t),Y(t)}^n \mid X(0) = i\big) \right\} = g_i^{s,n}(\pi^{-n}, R^n) \},
\end{aligned}$$

by which the theorem is proved. **QED**

Remark Theorem 2.8 with its conditions implies for all players the existence of stationary policies that are equilibrium policies, for any pay-off that is defined through limit-points (including Banach limits) of the sequence

$$\{\frac{1}{t+1} \mathbb{E}_R\big(\sum_{l=0}^{t} r_{X(l),Y(l)}^n \mid X(0) = i\big)\}_{t=0}^{\infty},$$

$n = 1, \ldots, N$, $i \in \mathsf{E}$.

3. α-DISCOUNT AND AVERAGE OPTIMALITY FOR 2-PERSON ZERO SUM GAMES

Fix again a discount factor $\alpha \in (0,1)$, and consider the case of two players. For simplicity we will adopt a different notation for the actions, policies, etc. of the two players, i.e. a, $A(i)$, π, Π will denote an action, the set of actions in state i, a stationary policy (randomised decision rule) and policy for player 1, and similarly b, $B(i)$, ρ and R for player 2.

Let the immediate rewards of player 2 be denoted by $r_{i,ab}$, $sup_{i,a,b}\mu_i^{-1}|r_{iab}| < \infty$, and let $V^\alpha(\Pi,\mathrm{R})$ denote her total expected α-discounted income for this reward function, when policy pair (Π,R) is used. $g(\Pi,\mathrm{R})$, $g^s(\Pi,\mathrm{R})$ then analogously denote the long-run expected average income and strong long-run average expected income of player 2. This game is simpler in the sense we have to deal with one value function only, but therefore also stronger results can be obtained.

The usual optimality criteria are the following. In the α-discounted game, Π^* is said to be an *optimal policy for player 1*, if

$$V_i^\alpha(\Pi^*,\mathrm{R}) \le \inf_\Pi \sup_\mathrm{R} V_i^\alpha(\Pi,\mathrm{R}), \quad \forall i \in \mathsf{E}, \forall \mathrm{R}.$$

This means that Π^* guarantees player 1 a maximum loss of $\inf_\Pi \sup_\mathrm{R} V_i^\alpha(\Pi,\mathrm{R})$.

Similarly, R^* is *optimal for player 2*, if

$$V_i^\alpha(\Pi,\mathrm{R}^*) \ge \sup_\mathrm{R} \inf_\Pi V_i^\alpha(\Pi,\mathrm{R}), \quad \forall i \in \mathsf{E}, \forall \Pi$$

and so R^* guarantees player 2 a minimum income of $\sup_\mathrm{R} \inf_\Pi V_i^\alpha(\Pi,\mathrm{R})$. If

$$\inf_\Pi \sup_\mathrm{R} V^\alpha(\Pi,\mathrm{R}) = \sup_\mathrm{R} \inf_\Pi V^\alpha(\Pi,\mathrm{R}),$$

then the game is said to have a value, and we will denote this value by V^α. Clearly the value is unique, provided it exists. The existence of a value does not necessarily imply the existence of an optimal strategy pair.

Replace $V^\alpha(\cdot,\cdot)$ by $g(\cdot,\cdot)$ to get the corresponding definitions for the average reward game.

The connection to the previous sections is given by the following lemma.

LEMMA 3.1 *Suppose that assumptions* 1^* *and* $2(\alpha)$ *hold,* $\alpha \in (0,1]$. *Then* (Π^*, R^*) *is a stationary optimal strategy pair for the two-person sum game if and only if* (Π^*, R^*) *is a stationary optimal strategy pair for the two-person non-cooperative game with*

$$r^1_{i,ab} = -r^2_{i,ab} = -r_{i,ab}.$$

Proof For $\alpha < 1$ this statement is trivial. We will only consider the case $\alpha = 1$, i.e. the average reward case. Let us assume that (Π^*, R^*) is a stationary optimal strategy pair for the two-person sum game. Then it immediately follows for the two-person non-cooperative game that (Π^*, R^*) is an AEP for player 2 and an s-AEP for player 1. But since strong average optimality is defined through the limsup and since for stationary policies the limsup and liminf are equal (see (2.13)), it follows that (Π^*, R^*) is also an AEP for the two-person non-cooperative game.

Conversely, assume that (Π^*, R^*) is a stationary optimal strategy pair for the two-person non-cooperative game. Then by theorem 2.8, (Π^*, R^*) is an s-AEP for player 1, and the result follows immediately. **QED**

3.1 Existence of α-discounted optimal policy pairs

The following result is a straightforward extension of the results in section 2.1.

THEOREM 3.2 *Suppose, that assumptions* 1^* *and* $2(\alpha)$ *hold. Then there exists a unique solution* $V^\alpha \in \mathsf{V}_\mu$ *to the optimality equation*

$$V_i^\alpha = \mathsf{val}\{r_{i,ab} + \alpha \sum_j P_{i\,ab\,j} V_j^\alpha\}. \tag{3.1}$$

V^α *is the value of the game. Any stationary policy pair* (π^*, ρ^*) *achieving the value in the righthandside of (3.1) is optimal and* $V^\alpha = V^\alpha(\pi^*, \rho^*)$.

Proof We first establish the unicity. Let v' and v'' be two solutions of (3.1). Then

$$v_i' - v_i'' \le \sup_{a,b} \sum_j \alpha P_{i\,ab\,j}(v_j' - v_j'').$$

Iterating this inequality we get the following (in the componentwize sense):

$$v' - v'' \le \sup_{\pi,\rho} \alpha^t P^t(\pi,\rho)(v' - v'') \le \sup_{\pi,\rho} \alpha^t P^t(\pi,\rho)|v' - v''|.$$

By symmetry, we may interchange between v' and v'' and obtain

$$|v' - v''| \le \sup_{\pi,\rho} \alpha^t P^t(\pi,\rho)|v' - v''|.$$

Taking the μ norm in both sides and choosing t sufficiently large, we see from (2.5) that $\|v' - v''\|_\mu \le c'\|v' - v''\|_\mu$ for some $c' < 1$, which implies that $v' = v$.

The other statements of the theorem follow from lemma 3.1 and theorem 2.3. **QED**

Next introduce the following notation: for $w \in \mathsf{V}_\mu$

$$L_\alpha(\pi,\rho)w = r(\pi,\rho) + \alpha P(\pi,\rho)w$$
$$\left(U_\alpha w\right)_i = \mathsf{val}\Big\{r_{i,ab} + \alpha \sum_j P_{i\,ab\,j} w_j\Big\}.$$

The following zero sum game version of the α-discounted *value iteration algorithm* converges:

- Select initial values $w_0^\alpha \in \mathsf{V}_\mu$.
- Define recursively $w_{t+1}^\alpha = U_\alpha w_t^\alpha$.
- Select (π_{t+1}, ρ_{t+1}) with
$$L_\alpha(\pi_{t+1}, \rho_{t+1}) w_t^\alpha = U_\alpha w_t^\alpha.$$

In order to prove convergence of this algorithm we will show that there exists T, such that U_α^T and $\prod_{t=1}^T L_\alpha(\pi_t, \rho_t)$ are contractions. Below we shall use the convention that $\max_{m' \in M} \mu_{m'} = 1$ if $M = \emptyset$.

LEMMA 3.3 *Suppose that assumptions* 1^* *and* $2(\alpha)$ *hold. Let* $\epsilon > 0$ *with* $\beta(1+\epsilon) < 1$ *and* T *such that*

$$\gamma = (1+\epsilon)^{-T}\left\{\frac{1}{1-\beta-\beta\epsilon} + \frac{\alpha+\alpha\epsilon}{(1-\alpha-\alpha\epsilon)(1-\beta-\beta\epsilon)}\max_{m'\in M}\mu_{m'}\right\} < 1.$$

Then for any $w, w' \in \mathsf{V}_\mu$

i)
$$\sup_{\left((\pi_1,\rho_1),\ldots,(\pi_T,\rho_T)\right)\in\mathcal{S}^T} \|L_\alpha(\pi_1,\rho_1)\cdots L_\alpha(\pi_T,\rho_T)(w-w')\|_\mu \le \gamma\|w-w'\|_\mu$$

ii)
$$\|U_\alpha^T w - U_\alpha^T w'\|_\mu \le \gamma\|w-w'\|_\mu$$

Proof As in the proof of lemma 2.1 it follows that for any T and for $\alpha' = \alpha(1+\epsilon)$, $\beta' = \beta(1+\epsilon)$, such that $\beta' < 1$

$$\|(P(\pi_1,\rho_1)\cdots P(\pi_T,\rho_T)\|_\mu \le (\alpha')^{-T}\left\{\frac{1}{1-\beta'} + \max_{m'\in M}\frac{\mu_{m'}\alpha'}{(1-\alpha')(1-\beta')}\right\}$$

for any arbitrary sequence (π_1,ρ_1), ..., (π_T,ρ_T). i) immediately follows. Next let (π_1,ρ_1),..., (π_T,ρ_T) and (π_1',ρ_1'), ..., (π_T',ρ_T') satisfy

$$\begin{aligned} U_\alpha^T w &= L_\alpha(\pi_1,\rho_1)\cdots L_\alpha(\pi_T,\rho_T)w \\ U_\alpha^T w' &= L_\alpha(\pi_1',\rho_1')\cdots L_\alpha(\pi_T',\rho_T')w'. \end{aligned}$$

Then
$$U_\alpha^T w - U_\alpha^T w' \le L_\alpha(\pi_1',\rho_1)\left(U_\alpha^{T-1}w - U_\alpha^{T-1}w'\right)$$
and by iterating this
$$U_\alpha^T w - U_\alpha^T w' \le L_\alpha(\pi_1',\rho_1)\cdots L_\alpha(\pi_T',\rho_T)(w-w').$$
Similarly
$$U_\alpha^T w - U_\alpha^T w' \ge L_\alpha(\pi_1,\rho_1')\cdots L_\alpha(\pi_T,\rho_T')(w-w')$$
and so ii) follows from i). **QED**

From (2.1) it follows that
$$\|U_\alpha^t w - U_\alpha^t w'\|_\mu \le \gamma^{\lfloor t/T\rfloor}c\|w-w'\|_\mu \tag{3.2}$$
for some constant $c > 0$, so that U_α has a fixed point in V_μ.

THEOREM 3.4 *Let assumptions* 1^* *and* $2(\alpha)$ *hold. Then* $\{w_t^\alpha\}$ *is a convergent sequence in* V_μ*. The limit* w_∞^α *satisfies the optimality equation (3.1) for any limit-point* (π_∞,ρ_∞) *of* $\{(\pi_t,\rho_t)\}$ *and hence* $w_\infty^\alpha = V^\alpha(\pi_\infty,\rho_\infty)$.

Proof It is clear that $\{w_t^\alpha\}$ converges to the unique fixed point, w_∞^α say, of the operator U_α. By the continuity of the value-operator it follows that w_∞^α is a solution to the optimality equation (3.1). It then easily follows that any limit-point of $\{(\pi_t,\rho_t)\}$ achieves the value in (3.1). **QED**

Because of U_α being a T-step contraction it is possible to derive similar bounds for the distance $\|w_t^\alpha - V^\alpha\|_\mu$ as in [44].

Finally, we note by the construction in Section 2.2, that w_t^α is the value of a Markov game with horizon t, and that the Markov policies $\big((\pi_t,\rho_t(,(\pi_{t-1},\rho_{t-1},...,(\pi_0,\rho_0)\big)$ are optimal for that problem.

3.2 Existence of average optimal policy pairs

The derivations in this case are more complicated and we have to use the strong optimality result theorem 2.8 (which was used for establishing lemma 3.1 for the average reward case). Consider the following pair of functional equations (cf. [16]):

$$g_i = \mathsf{val}\Big\{\sum_j P_{i\,ab\,j} g_j\Big\}, \qquad i\in \mathsf{E} \tag{3.3a}$$

$$v_i = \mathsf{val}_{A(i,g),B(i,g)}\Big\{r_{iab} - g_i + \sum_j P_{i\,ab\,j} v_j\Big\}, \quad i\in \mathsf{E}, \tag{3.3b}$$

where $A(i,g)$ and $B(i,g)$ are the sets of optimal randomized actions in (3.3a) for the vector g. In terms of the value operator a generalisation of the weaker functional equations (2.9a,b,c) does not seem possible. Obviously the assertion in theorem 2.6i) holds for $v^1=-v^2$ and $g^1=-g^2$. We have the following result.

THEOREM 3.5 *Suppose, that assumptions* 1^* *and* $2(1)$ *hold and that* ν *is continuous on* $\mathcal{S}$.

i) Suppose there exists a solution pair $(g,v)\in(\mathsf{V}_\mu)^2$ *to the optimality equations (3.3). Then* g *is the unique value of the game. Any stationary policy pair* (π^*,ρ^*) *achieving the value in the righthandside of (3.3a,b) is optimal.*

ii) For any limit-point of α*-discounted optimal policy pairs,* $\alpha\to 1$*, the pair* (g,v) *of the form*

$$g = \Pi(\pi,\rho) r(\pi,\rho)$$
$$v = D(\pi,\rho) r(\rho,\pi).$$

is a solution to (3.3).

Proof It is worthwhile to give some details. Consider a solution $(g,v)\in(\mathsf{V}_\mu)^2$ to (3.3). It can be shown that the sets $A(i,g)$ and $B(i,g)$ are compact (by μ-continuity) and convex, and thus there exists a policy pair, (π,ρ) say, achieving the value in (3.3) (see the section on lexicographical games in [4]). By writing $r^1_{iab}=-r_{iab}$ and $r^2_{iab}=r_{iab}$ this can be transformed into a solution $(\hat g,\hat v)\in(\mathsf{V}_\mu)^4$ to the optimality equations (2.9). For this solution we have $\hat g^1=-\hat g^2=-g$ and $\hat v^1=-\hat v^2=-v$. It directly follows that $g=\hat g^2\ge g(\pi,\mathrm{R})$ for any policy R of player 2. From the strong optimality result theorem 2.8 it follows that

$$\hat g^2_i = -\hat g^1_i \le -g_i^{s,1}(\Pi,\hat\rho) = \liminf -\frac{1}{t+1}\mathbb{E}_{(\Pi,\hat\rho)}\Big(\sum_{k=0}^t r_{X(t),Y(t)}\mid X(0)=i\Big) = g_i^2(\Pi,\hat\rho),$$

for any policy Π for player 1. Hence g is the unique value of the game and any policy pair achieving the value in (3.3) is optimal. This proves i). ii) follows from theorem 2.6 by converting the explicit solution given there to the required solution of (3.3). **QED**

It is simple to prove the obvious extension of lemma 2.7: under assumptions 1^* and $2(1)$ there exists a solution $(g,\hat v)$ to the functional equations

$$g_i = \mathsf{val}\Big\{\sum_j P_{i\,ab\,j} g_j\Big\}, \quad i\in\mathsf{E}$$

$$\hat v_i = \mathsf{val}\Big\{r_{iab} - g_i + \sum_j P_{i\,ab\,j}\hat v_j\Big\}, \quad i\in\mathsf{E}.$$

Similar as in theorem 2.8 this implies the existence of a "strong" average optimal policy pair.

For $\nu = 1$ theorem 3.5 can be simplified.

THEOREM 3.6 *Under the conditions of theorem 3.5 the following holds for $\nu = 1$: there is a solution pair (g, v) to the optimality equation*

$$v_i = \mathsf{val}\{r_{iab} - g + \sum_l P_{i\,ab\,j} v_j\}, \; i \in \mathsf{E}. \tag{3.4}$$

g is a constant and it is the unique value of the game. $v \in \mathsf{V}_\mu$ and it is unique upto a constant (within the space V_μ). Any policy pair achieving the value in (3.4) is optimal.

Proof The only thing left to prove is the unicity of v. Let $(g, v + d)$ be another solution of (3.4), with $d \in \mathsf{V}_\mu$. Let (π, ρ) and $(\hat{\pi}, \hat{\rho})$ be policy pairs achieving the value in (3.4) for (g, v) and $(g, v + d)$ respectively. Then $d \geq P(\hat{\pi}, \rho)d$. Select a state $b \in M$ that is positive recurrent in the MC induced by policy $(\hat{\pi}, \rho)$ and define $\hat{d}_i = d_i - d_b$. Then

$$\hat{d} \geq {}_bP(\hat{\pi}, \rho)\hat{d}.$$

Iterating yields

$$\hat{d} \geq {}_bP^n(\hat{\pi}, \rho)\hat{d},$$

the righthandside of which converges to 0 for $n \to \infty$. Thus $d_i \geq d_b$. By iterating $d \geq P(\hat{\pi}, \rho)d$ we find

$$d_b \geq \sum_j \Pi_{bj}(\hat{\pi}, \rho) d_j$$

and it follows that $d_i = d_b$ for states i that are positive recurrent under $(\hat{\pi}, \rho)$.

An analogous reasoning gives that $d_i \leq d_{b'}$ for $b' \in M$ a positive recurrent state under policy $(\pi, \hat{\rho})$ and $d_i = d_{b'}$ for all positive recurrent states i under this policy. Since $\nu = 1$ for all stationary policies, the sets of positive recurrent states under $(\hat{\pi}, \rho)$ and $(\pi, \hat{\rho})$ have a non-empty intersection, containing the state i_0 say. Consequently

$$d_i \geq d_b = d_{i_0} = d_{b'} \geq d_i,$$

so that $d_i = d_{i_0}$ for all i. **QED**

A remark is due. If the α-discounted rewards have a Puiseux series expansion in $\alpha = 1$, then under our assumptions it follows that it is of the following form

$$V^\alpha = \frac{g}{\rho} + \sum_{k \geq 0} x_{k/K} \rho^{k/K},$$

for $\rho = (1 - \alpha)/\alpha$ the interest rate and K some integer. That is: the Puiseux series has only positive fractional powers. It is an open problem whether the whole Puiseux series expansion exists and if so, how to compute K.

We will next consider the *value iteration* algorithm:

- Select an initial vector $w_0 \in \mathsf{V}_\mu$.
- Define recursively $w_{t+1} = U_1 w_t$ (i.e. $w_{i,t+1} = \mathsf{val}\Big\{r_{iab} + \sum_j P_{i\,ab\,j} w_{j,t}\Big\}$)

- Select (π_{t+1}, ρ_{t+1}) with $L_1(\pi_{t+1}, \rho_{t+1}) = U_1 w_t$.

Further let (g, v) a solution to (3.3) be given. Let (π^*, ρ^*) be a policy pair achieving the value in (3.3) and define

$$e_t = w_t - tg - v.$$

So far we have only been able to prove convergence of the value iteration algorithm under a strong unichain condition in conjunction with μ-uniform geometric recurrence (assumption 2(1)). This condition will be satisfied in most applications. To ensure uniform aperiodicity we will also assume a uniform lower bound on the probability of staying in the same state. This is no restriction. We can use the data-transformation (cf. [38])

$$\hat{P}_{i\,ab\,j} = \epsilon\delta_{ij} + (1-\epsilon)P_{i\,ab\,j},$$

for which stationary optimal policies in the original problem are optimal in the transformed problem and vice versa. Note that assumption 2(1) holds for the transformed problem with a different contraction rate.

Assumption 3

i) *Strong unichain condition*: There exists is a state, 0 say, that is positive recurrent in the MC induced by any stationary strategy pair.

ii) There exists $\epsilon > 0$, such that $P_{i\,ab\,i} \geq \epsilon$ for $(a, b) \in A(i) \times B(i)$, $i \in \mathsf{E}$.

By assumption 3i) we only need to consider one optimality equation, instead of a set of two equations. In the case of one player we do not need this extra condition, because then optimality is a global criterion and not a local criterion as is the case with two players. As has been said before, assumption 3ii) is no restriction, and we need it only for proving convergence of $\{e_t\}$. Our analysis is based on [20] and [41], although the proof is more complicated.

We will first derive some useful bounds.

Lemma 3.7 *Under assumptions* 1^*, 2(1) *and 3(i) the following hold:*

i) $P(\pi^*, \rho_t)e_{t-1} \geq e_t \geq P(\pi_t, \rho^*)e_{t-1}$.

ii) $\sup_t \|e_t\|_\mu < \infty$.

iii) $\lim_{t\to\infty} \|g - \frac{1}{t}w_t\|_\mu = 0$.

Remark that the third assertion states that g is the asymptotic average value of the maximum expected reward for player 2 over t time periods. This result will be sharpened considerably in theorem 3.7 by showing that $w_t - tg$ converges as $t \to \infty$.

Proof First note that $\|e_0\|_\mu = \|w_0 - v\|_\mu \leq \|w_0\|_\mu + \|v\|_\mu < \infty$. Further

$$\begin{aligned}
P(\pi^*, \rho_t)e_{t-1} &= P(\pi^*, \rho_t)\{w_{t-1} - (t-1)g - v\} \\
&= r(\pi^*, \rho_t) + P(\pi^*, \rho_t)w_{t-1} - tg - \big(r(\pi^*, \rho_t) - g + P(\pi^*, \rho_t)v\big) \\
&\geq w_t - tg - v = e_t.
\end{aligned}$$

The second inequality in i) is proved similarly. For the proof of ii) note that

$$e_t \geq P(\pi^*, \rho_t)e_{t-1} \geq \cdots \geq P(\pi^*, \rho_t)\cdots P(\pi^*, \rho_1)e_0.$$

As in inequality (2.15) there exists a constant c' such that

$$e_{i,t} \geq -\sum_j \Big[P(\pi^*,\rho_t)\cdots P(\pi^*,\rho_1)\Big]_{ij}\mu_j \frac{|e_{j,0}|}{\mu_j} \geq -c'\|e_0\|_\mu\ \mu_i, \quad t \geq 0,$$

and similarly $e_{i,t} \leq c'\|e_0\|_\mu\ \mu_i$, $t \geq 0$. This shows ii). Finally $g = \frac{1}{t}\big(w_t - e_t - v\big)$. By ii)

$$\|g - \frac{1}{t}w_t\|_\mu\ = \frac{1}{t}\|e_t + v\|_\mu\ \leq \frac{1}{t}\big(c'\|e_0\|_\mu\ + \|v\|_\mu\ \big) \to 0, \quad t \to \infty. \qquad \textbf{QED}$$

This enables us to show the main theorem.

THEOREM 3.8 *Under assumptions* 1^*, $2(1)$ *and* 3 *the following hold.*
i) $\{e_t\}_t$ *is a convergent sequence with limit, say* d, *a constant vector.*
ii) Any vector limit-point of the sequence $\{\pi_t,\rho_t\}_t$ *is an optimal policy pair achieving the value in (3.4).*

Proof From Key theorem I in [12] we may assume without loss of generality that $\sum_j {}_0P_{i\,ab\,j}\mu_j \leq \beta\mu_i$ for some $\beta < 1$. We first show convergence of the sequence $\{e_t\}_t$. Write

$$\begin{aligned} m_i &= \liminf_{t\to\infty} e_{i,t} \\ M_i &= \limsup_{t\to\infty} e_{i,t} \\ d_{iab} &= r_{iab} - g + \sum_j P_{i\,ab\,j} v_j - v_i. \end{aligned}$$

Then

$$\begin{aligned} e_{i,t+1} &= \mathsf{val}\{r_{iab} + \sum_j P_{i\,ab\,j} w_{j,t} - (t+1)g - v_i\} \\ &= \mathsf{val}\{r_{iab} + \sum_j P_{i\,ab\,j}\big(w_{j,t} - tg - v_j\big) - g + \sum_j P_{i\,ab\,j} v_j - v_i\} \\ &= \mathsf{val}\{d_{iab} + \sum_j P_{i\,ab\,j} e_{j,t}\}. \end{aligned} \tag{3.5}$$

Fix $i \in \mathsf{E}$ and let $\{t(k)\}_k$ be a sequence with $e_{i,t(k)} \to M_i$, $k \to \infty$; $e_{j,t(k)-1} \to \hat{e}_j$, $j \in \mathsf{E}$, for some vector $\hat{e} \in \mathsf{V}_\mu$ and $\rho_{i,t(k)} \to \hat{\rho}_i$. It is always possible to achieve this by the μ-boundedness of $\{e_t\}_t$, by diagonalisation and by the compactness of $\mathcal{S}^2(i)$. By continuity and monotonicity of the value operator

$$M_i = \mathsf{val}\{d_{iab} + \sum_j P_{i\,ab\,j}\hat{e}_j\} \leq \mathsf{val}\{d_{iab} + \sum_j P_{i\,ab\,j} M_j\} \tag{3.6}$$

and further by lemma 3.7

$$M_i \leq \sum_j P_{ij}(\pi^*,\hat{\rho}) M_j. \tag{3.7}$$

Similarly

$$m_i \geq \mathsf{val}\{d_{iab} + \sum_j P_{i\,ab\,j} m_i\} \tag{3.8}$$

and

$$m_i \geq \sum_j P_{ij}(\hat{\pi}, \rho^*) m_j, \tag{3.9}$$

for some limit-point $\hat{\pi}$ of $\{\pi_t\}_t$. Observe that $\|M\|_\mu$, $\|m\|_\mu$ $< \infty$. Note that the necessary interchanges of limits and summations are allowed due to μ-boundedness and μ-continuity properties. Let $\widehat{M}_i = M_i - M_0$, then (3.7) implies

$$\widehat{M}_i \leq \sum_j {}_0P_{ij}(\pi^*, \hat{\rho})\widehat{M}_j.$$

Iterate this inequality:

$$\widehat{M} \leq {}_0P^t(\pi^*, \hat{\rho})\widehat{M} \leq \beta^t \|\widehat{M}\|_\mu \, \mu, \quad t \geq 0.$$

Letting $t \to \infty$ gives $\widehat{M} \leq 0$, i.e. $M_i \leq M_0$. Iterating (3.7) also yields for $i = 0$

$$M_0 \leq \sum_j \Pi_{0j}(\pi^*, \hat{\rho}) M_j,$$

so that $M_i = M_0$ for $i \in R(\pi^*, \hat{\rho})$, the set of positive recurrent states under $(\pi^*, \hat{\rho})$. Similarly $m_i \geq m_0$ and $m_i = m_0$ for $i \in R(\hat{\pi}, \rho^*)$, the set of positive recurrent states under policy $(\hat{\pi}, \rho^*)$. Hence $0 \leq M_i - m_i \leq M_0 - m_0$ and $\{e_t\}_t$ converges, if we can show that $M_0 = m_0$.

Consider sequences $\{t(k)\}_k$, $\{m(k)\}_k$ with

$$\begin{aligned} e_{0,t(k)} &\to M_0 \\ e_{0,m(k)} &\to m_0 \\ m(k) - t(k) &\to \infty, \end{aligned}$$

for $k \to \infty$. As a shorthand notation let (t, t') denote the policy $\big((\pi_{t-1}, \rho^*), \ldots, (\pi_{t'}, \rho^*)\big)$, which is played for $t - t'$ time instants and write

$$p_{j\bullet}(t, t') = \Big[P(\pi_{t-1}, \rho_{t-1}) \cdots P(\pi_{t'}, \rho_{t'})\Big]_{j\bullet}.$$

By uniform integrability under μ-uniform geometric recurrence (cf. lemma 4.7, [12]), the sequence of probability distributions $\{p_{0\bullet}\big(m(k), t(k)\big)\}_k$ is tight and so it has a convergent subsequence, with limit a probability distribution, p say. It follows that

$$\begin{aligned} m_0 = \lim_k e_{0,m(k)} &\geq \lim_k \sum_j p_{0j}\big(m(k), t(k)\big) e_{j,t(k)} \\ &\geq \sum_j p_j \liminf_k e_{j,t(k)} \\ &\geq p_0 M_0 + (1 - p_0) m_0, \end{aligned}$$

by Fatou's lemma. This implies $m_0 = M_0$, if we can show that $p_0 > 0$.

From lemma 4.7, [12] and μ-uniform geometric recurrence, there exists a finite set K and $\delta \in (0,1)$, such that for all policy pairs (Π, R)

$$\begin{cases} \displaystyle\sum_{j \notin K} \mathbb{P}_{(\Pi,\mathrm{R})}\{X(t) = j \mid X(0) = 0\}\mu_j \leq \delta, & t \geq 0 \\ \mathbb{P}_{(\Pi,\mathrm{R})}\{\tau_0 > t | X(0) = i\} \leq \delta, & i \in K, t \geq T, \end{cases}$$

for some $T > 0$ and with τ_0 the first hitting time of state 0. The first inequality clearly implies

$$\mathbb{P}_{(\Pi,\mathrm{R})}\{X(t) \in K \mid X(0) = 0\} \geq 1 - \delta, \quad t \geq 0, \tag{3.10}$$

and the second that

$$\mathbb{P}_{(\Pi,\mathrm{R})}\{\tau_0 \leq t | X(0) = i\} \geq 1 - \delta, \quad i \in K, t \geq T. \tag{3.11}$$

Then for $T' > T$ fixed and noting that $T' < m(k) - t(k)$ for k sufficiently large

$$\begin{aligned} p_{00}\big(m(k), t(k)\big) &= \sum_j p_{0j}\big(m(k), t(k) + T'\big) \cdot p_{j0}\big(t(k) + T', t(k)\big) \\ &\geq \sum_{j \in K} p_{0j}\big(m(k), t(k) + T'\big) \sum_{r=0}^{T'-1} \mathbb{P}_{(t(k)+T', t(k))}\{\tau_0 = T' - 1 - r | X(0) = j\}\epsilon^r \\ &\geq \sum_{j \in K} p_{0j}\big(m(k), t(k) + T'\big) \mathbb{P}_{(t(k)+T', t(k))}\{\tau_0 \leq T' - 1 | X(0) = j\}\epsilon^{T'} \\ &\geq (1 - \delta)^2 \epsilon^{T'}. \end{aligned}$$

by (3.10), (3.11) and assumption 3ii) and so $p_0 > 0$ and so $m_0 = M_0$.

Next combine (3.6) and (3.8) and the fact that $M_i = m_i$ to obtain

$$v_i + m_i = \mathsf{val}\{r_{iab} - g + \sum_j P_{i\,ab\,j}(v_j + m_j)\}$$

and hence by theorem 3.6 $m_i = m_0$, so that the sequence $\{e_t\}$ converges to a vector with constant components.

Finally note that (π_{t+1}, ρ_{t+1}) achieve the value in (3.5). From the convergence of $\{e_t\}$ and the continuity of the value operator it follows that any limit point of $\{(\pi_t, \rho_t)\}$ achieves the value in (3.4).

QED

4. Control of birth-death processes: an example.

The results in the previous section will be applied to the following model. It is a stochastic game on the state space $\mathsf{E} = \mathbb{Z}_+$, where player 1 controls the death rates $\chi \in [\underline{\chi}, \overline{\chi}]$ and independently player 2 controls the birth rates $\lambda \in [\underline{\lambda}, \overline{\lambda}]$. When the state is i and action (χ_i, λ_i) is chosen, player 1 has to pay to player 2 an amount $c_i(\chi_i, \lambda_i)$ per unit time. In other words, we consider the control of a continuous time Markov jump process with a countable state space E, where the transition rates from state i to $i+1$ are controlled by player 1, and the transition rates from state i to $i-1$ (for $i > 0$) are controlled by player 2.

Instead of analysing directly the continuous time model, we shall study the corresponding standard time-discretised version. Under fairly general assumptions on the cost functions, we will derive optimality of deterministic stationary policies within the class of all policies for the time-discretised model.

It can now be shown, using standard uniformisation techniques (see Lippman [24] and Serfozo [40]), that the equilibrium stationary policies for the time-discretised model are also equilibrium policies for the original continuous time problem, if one restricts in the continuous model to policies where the control actions may change only at jump epochs of the state process.

The transition probabilities for the discretised model are as follows. For $(\chi,\lambda)\in[\underline{\chi},\overline{\chi}]\times[\underline{\lambda},\overline{\lambda}]$

$$P_{0j}=\begin{cases}1-\lambda, & j=0\\ \lambda, & j=1\end{cases}\qquad P_{ij}=\begin{cases}\lambda, & j=i+1\\ \chi, & j=i-1\\ 1-\lambda-\chi, & j=i\end{cases}$$

by suitably renormalising. We assume the costs to be separable:

$$c_i(\chi,\lambda)=h(i)+\zeta(\lambda)+\xi(\chi),$$

with ζ and ξ continuous functions. It is the objective to find a Nash-equilibrium pair of policies for α-discounted and average optimality criteria.

In applications, it is natural to assume that $\zeta(\lambda)$ is non-increasing and $\xi(\chi)$ is non-decreasing. This assumption is not required, however, for the results to hold.

To satisfy assumptions 1^*, 3 and $2(\alpha)$, we have to make the following assumptions.

Assumption $4(\alpha)$: h is convex non-decreasing, such that for some x, $1<x<\alpha^{-1}$,

$$\sup_{i\in\mathsf{E}} h(i)x^{-i}<\infty.$$

It is easy to check that assumption $4(\alpha)$ implies assumptions 1^* and $2(\alpha)$ for a bounding vector μ given by $\mu_i=x^i$ and $M=\emptyset$. Indeed, assumption 1^* is a direct consequence of assumption $4(\alpha)$ and assumption $2(\alpha)$ follows since for $\beta=\alpha x$, we have

$$\alpha\sum_j {}_MP_{iabj}\mu_j\le\alpha\sum_j P_{iabj}x^j\le\alpha x^{i+1}=\beta x^i=\beta\mu_i,\quad\forall a,b,i\ .$$

As a consequence the results of theorems 3.2, 3.4, and 3.6 hold: there exist stationary optimal policy pairs and the value iteration algorithm converges for the α-discounted criteria.

As an application of such a model one might think of a service centre having a single queue. The service capacity is controlled, whilst the arrival rates are only known to vary between two parameters. It is the objective to design a minimum cost policy under worst case arrival conditions. It is natural to have a cost function consisting of: holding costs for having a high number of customers in the system, costs for having a large service capacity, and a reward for having a higher number of customers entering the centre. This is reflected in our conditions on the cost function.

Theorem 4.1 *Suppose, that assumption $4(\alpha)$ holds.*

i) Then there are positive constants $\bar{\imath}$ and $\alpha^0<1$, such that for $\alpha\ge\alpha^0$ there exists an optimal policy pair $(\chi^\alpha,\lambda^\alpha)$ with

$$\lambda_i^\alpha=\overline{\lambda},\quad\chi_i^\alpha=\overline{\chi},\quad i\ge\bar{\imath},$$

for $(\chi_i^\alpha,\lambda_i^\alpha)$ the decision in state i. Moreover, λ_i^α and χ_i^α are non-decreasing in i.

ii) If in addition $\zeta(\lambda)$ is convex and $\xi(\chi)$ concave then $(\chi_\alpha,\lambda_\alpha)$ is "bang-bang", i.e. there are constants i^0, i^1 with

$$\lambda_i^\alpha=\begin{cases}\underline{\lambda}, & i<i^0\\ \overline{\lambda}, & i\ge i^0\end{cases},\qquad\chi_i^\alpha=\begin{cases}\underline{\chi}, & i<i^1\\ \overline{\chi}, & i\ge i^1.\end{cases}$$

For the proof we use value-iteration. Let $w_0^\alpha = 0$ and w_t^α be the optimal value vector over t time periods. We shall understand throughout that $w_{-i,t}^\alpha = w_{0,t}^\alpha$ for $i > 0$. Then

$$\begin{aligned}
w_{i,t+1}^\alpha &= \inf_\chi \sup_\lambda \left[h(i) + \zeta(\lambda) + \xi(\chi) + \alpha\{\lambda w_{i+1,t}^\alpha + \chi w_{i-1,t}^\alpha + (1-\lambda-\chi) w_{i,t}^\alpha\}\right] \\
&= h(i) + \alpha w_{i,t}^\alpha + \sup_\lambda \left[\zeta(\lambda) + \alpha\lambda\{w_{i+1,t}^\alpha - w_{i,t}^\alpha\}\right] + \inf_\chi \left[\xi(\chi) - \alpha\chi\{w_{i,t}^\alpha - w_{i-1,t}^\alpha\}\right] \\
&= h(i) + \alpha w_{i,t}^\alpha + \sup_\lambda s_{i,t}^\alpha(\lambda) + \inf_\chi r_{i,t}^\alpha(\chi), \qquad (4.1)
\end{aligned}$$

with

$$\begin{aligned}
r_{i,t}^\alpha(\chi) &= \xi(\chi) - \alpha\chi\{w_{i,t}^\alpha - w_{i-1,t}^\alpha\}, \\
s_{i,t}^\alpha(\lambda) &= \zeta(\lambda) + \alpha\lambda\{w_{i+1,t}^\alpha - w_{i,t}^\alpha\}.
\end{aligned}$$

Then the policy pair $(\chi_t^\alpha, \lambda_t^\alpha)$ given by

$$\begin{aligned}
\chi_{i,t}^\alpha &= \inf\{\chi \mid r_{i,t-1}^\alpha(\chi) = \inf_\chi r_{i,t-1}^\alpha(\chi)\} \\
\lambda_{i,t}^\alpha &= \inf\{\lambda \mid s_{i,t-1}^\alpha(\lambda) = \sup_\lambda s_{i,t-1}^\alpha(\lambda)\}
\end{aligned}$$

is optimal in the t^{th} iteration. In order to complete the proof, we need the following lemmas.

LEMMA 4.2 *If h is convex non-decreasing, then for all t,*
i) $\lambda_{i,t}^\alpha$ *is monotone non-decreasing in* i
ii) $\chi_{i,t}^\alpha$ *is monotone non-decreasing in* i
iii) $w_{i,t}^\alpha$ *is convex non-decreasing in* i
iv) a. $\zeta(\lambda)$ *convex implies that* $\lambda_{i,t}^\alpha \in \{\underline{\lambda}, \overline{\lambda}\}$ *for all* i
b. $\xi(\chi)$ *concave implies that* $\chi_{i,t}^\alpha \in \{\underline{\chi}, \overline{\chi}\}$ *for all* i.

Proof For $t = 1$ the assertions are trivial. Assume that the assertions hold upto $t-1$. First suppose, that i) is violated for t. Then there are states $i_1 < i_2$ with $\lambda_{i_1,t}^\alpha > \lambda_{i_2,t}^\alpha$. Write $\tilde\lambda_k = \lambda_{i_k,t}^\alpha$, then by the optimality of $\tilde\lambda_k$ in state i_k, $k = 1, 2$, at time t,

$$\begin{aligned}
s_{i_1,t-1}^\alpha(\tilde\lambda_1) &> s_{i_1,t-1}^\alpha(\tilde\lambda_2) \\
s_{i_2,t-1}^\alpha(\tilde\lambda_1) &\le s_{i_2,t-1}^\alpha(\tilde\lambda_2).
\end{aligned}$$

Hence

$$s_{i_1,t-1}^\alpha(\tilde\lambda_1) - s_{i_1,t-1}^\alpha(\tilde\lambda_2) > s_{i_2,t-1}^\alpha(\tilde\lambda_1) - s_{i_2,t-1}^\alpha(\tilde\lambda_2). \qquad (4.2)$$

The induction hypothesis implies for $\lambda_1 > \lambda_2$ that

$$s_{i,t-1}^\alpha(\lambda_1) - s_{i,t-1}^\alpha(\lambda_2) = \zeta(\lambda_1) - \zeta(\lambda_2) + \alpha(\lambda_1 - \lambda_2)\{w_{i+1,t-1}^\alpha - w_{i,t-1}^\alpha\}$$

is non-decreasing in i, contradicting (4.2). This proves i) for t. The proof of ii) for t is analogous. iv) follows directly from (4.1) and the monotonicity of $w_{i,t-1}^\alpha$ in i (induction hypothesis). We prove iii) for t. Define

$$g_i = \alpha w_{i,t-1}^\alpha + \inf_\chi r_{i,t-1}^\alpha(\chi) + \sup_\lambda s_{i,t-1}^\alpha(\lambda),$$

and

$$\Delta_i = g_{i+1} - g_i - [g_i - g_{i-1}].$$

In order to prove the convexity of $w^\alpha_{i,t}$, it is sufficient to show that g_i is convex, or, more precisely, that $\Delta_i \geq 0$ for all $i > 0$. Since $g_0 = g_i$ for $i < 0$, (by definition of $w^\alpha_{i,t}$ for negative i), it is sufficient to show that $\Delta_0 \geq 0$ in order to conclude that $g_1 \geq g_0$, which, together with the convexity of g_i, implies that g_i is monotone non-decreasing. Thus, by establishing that $\Delta_i \geq 0$ for all $i \geq 0$ we conclude that $w^\alpha_{i,t}$ are monotone non-decreasing and convex.

By the induction hypothesis iii) for w^α_{t-1} we have for $i \geq 0$:

$$\begin{aligned}
\Delta_i \geq & \alpha \left(w^\alpha_{i+1,t-1} - w^\alpha_{i,t-1} - w^\alpha_{i,t-1} + w^\alpha_{i-1,t-1}\right) \\
& + s^\alpha_{i+1,t-1}(\lambda^\alpha_{i,t}) - s^\alpha_{i,t-1}(\lambda^\alpha_{i,t}) - [s^\alpha_{i,t-1}(\lambda^\alpha_{i,t}) - s^\alpha_{i-1,t-1}(\lambda^\alpha_{i,t})] \\
& + r^\alpha_{i+1,t-1}(\chi^\alpha_{i+1,t}) - r^\alpha_{i,t-1}(\chi^\alpha_{i+1,t}) - [r^\alpha_{i,t-1}(\chi^\alpha_{i-1,t}) - r^\alpha_{i-1,t-1}(\chi^\alpha_{i-1,t})] \\
= & \alpha \left(w^\alpha_{i+1,t-1} - w^\alpha_{i,t-1} - w^\alpha_{i,t-1} + w^\alpha_{i-1,t-1}\right) \\
& - \alpha\chi^\alpha_{i+1,t} \left(w^\alpha_{i+1,t-1} - w^\alpha_{i,t-1} - w^\alpha_{i,t-1} + w^\alpha_{i-1,t-1}\right) \\
& + \alpha\chi^\alpha_{i-1,t} \left(w^\alpha_{i,t-1} - w^\alpha_{i-1,t-1} - w^\alpha_{i-1,t-1} + w^\alpha_{i-2,t-1}\right) \\
& + \alpha\lambda^\alpha_{i,t} \left(w^\alpha_{i+2,t-1} - w^\alpha_{i+1,t-1} - w^\alpha_{i+1,t-1} + w^\alpha_{i,t-1} - [w^\alpha_{i+1,t-1} - w^\alpha_{i,t-1} - w^\alpha_{i,t-1} + w^\alpha_{i-1,t-1}]\right) \\
= & \alpha(1 - \lambda^\alpha_{i,t} - \chi^\alpha_{i+1,t}) \left(w^\alpha_{i+1,t-1} - w^\alpha_{i,t-1} - w^\alpha_{i,t-1} + w^\alpha_{i-1,t-1}\right) \\
& + \alpha\lambda^\alpha_{i,t} \left(w^\alpha_{i+2,t-1} - w^\alpha_{i+1,t-1} - w^\alpha_{i+1,t-1} + w^\alpha_{i,t-1}\right) \\
& + \alpha\chi^\alpha_{i-1,t} \left(w^\alpha_{i,t-1} - w^\alpha_{i-1,t-1} - w^\alpha_{i-1,t-1} + w^\alpha_{i-2,t-1}\right) \\
\geq & 0.
\end{aligned}$$

QED

By virtue of theorem 3.4 we conclude the following.

COROLLARY 4.3 *Under assumption* 4(α) *the optimal value function* $V^\alpha = \lim_{t\to\infty} w^\alpha_t$ *is convex non-decreasing, and it satisfies the optimality equation (3.1) for any limit-point* $(\chi^\alpha, \lambda^\alpha)$ *of the sequence* $\{(\chi^\alpha_t, \lambda^\alpha_t)\}$. $(\chi^\alpha, \lambda^\alpha)$ *is monotone non-decreasing in* λ *and* χ *as a function of the state variable* i. *In particular, if* $\zeta(\lambda)$ *is convex and* $\xi(\chi)$ *concave, then* $(\chi^\alpha, \lambda^\alpha)$ *is "bang-bang".*

We will use this to show insensitivity of the optimal policy pair in the t^{th} iteration of the value algorithm for large states and α close to 1. The next lemma uses arguments from [14, 22]. For any real function f on E, define $\Delta f_i = f_{i+1} - f_i$.

LEMMA 4.4 *Suppose that that there exists a constant* $c > 0$, *such that* $\Delta h(i) \geq c$ *for all* i, *then*
i) $\lim_{a\uparrow 1} w^\alpha_{i+1,t} - w^\alpha_{i,t} \geq c \cdot \min\{i, t\}$
ii) $\lim_{\alpha\uparrow 1} \lim_{i\to\infty} \lim_{t\to\infty}(w^\alpha_{i+1,t} - w^\alpha_{i,t}) = \infty$.

This condition on h is no restriction, provided we require h to be convex strictly increasing for i large enough.

Proof From (4.1) we obtain for $i, t \geq 1$

$$\begin{aligned}
\Delta w^\alpha_{i,t} & \geq c + \Delta s^\alpha_{i,t-1}(\lambda^\alpha_{i,t-1}) + \Delta r^\alpha_{i,t-1}(\chi^\alpha_{i+1,t-1}) + \alpha\Delta w^\alpha_{i,t-1} \\
& = c + \alpha\lambda^\alpha_{i,t-1}\{\Delta w^\alpha_{i+1,t-1} - \Delta w^\alpha_{i,t-1}\} - \alpha\chi^\alpha_{i+1,t-1}\{\Delta w^\alpha_{i,t-1} - \Delta w^\alpha_{i-1,t-1}\} + \alpha\Delta w^\alpha_{i,t-1} \\
& = c + \alpha\lambda^\alpha_{i,t-1}\{\Delta w^\alpha_{i+1,t-1} - \Delta w^\alpha_{i,t-1}\} + \alpha(1 - \chi^\alpha_{i+1,t-1})\{\Delta w^\alpha_{i,t-1} - \Delta w^\alpha_{i-1,t-1}\} + \alpha\Delta w^\alpha_{i-1,t-1} \\
& \geq c + \alpha\Delta w^\alpha_{i-1,t-1},
\end{aligned}$$

and so

$$\Delta w_{i,t}^{\alpha} \geq c\left(1 + \alpha + \cdots + \alpha^{\min\{i,t\}}\right).$$

The result follows. **QED**

The assertions in theorem 4.1 follow immediately from theorem 3.4. **QED**

Next we consider the expected average cost, and make the following assumption:

ASSUMPTION 5:

i) *Weak stability*

$$\overline{\lambda} < \overline{\chi}.$$

ii) h is convex and strictly increasing at at most a polynomial rate, i.e. for all x, $1 < x < \overline{\chi}/\overline{\lambda}$,

$$\sup_{i \in \mathsf{E}} h(i) x^{-i} < \infty.$$

THEOREM 4.5 *The results of theorem 4.1 hold for the expected average cost with assumption 5 replacing assumption* $4(\alpha)$.

Proof We first note that under the stronger stability assumption

$$\overline{\lambda} < \underline{\chi},$$

the results follow from theorem 3.5. Indeed, the strong stability assumption implies that assumptions 1^* and 2(1) hold, see [23], by choosing $\mu_i = y^i$ where y is any number satisfying $1 < y < \underline{\chi}/\overline{\lambda}$.

The weak stability is more involved. Let us note, that assumption 5 implies assumption $4(\alpha)$ for all α sufficiently close to 1. Let $\alpha(n)$ be a sequence of discount factors, converging to one. Let $\{\chi^{\alpha(n)}, \lambda^{\alpha(n)}\}_n$ be a sequence of pure stationary policies for both players, such that $\chi^{\alpha(n)}$ and $\lambda^{\alpha(n)}$ are optimal for $\alpha(n)$, and satisfy the structure given in theorem 4.1. Consider any pure stationary policies χ^1 and λ^1 for both players obtained as accumulation points of these sequences. They clearly inherit the structure of the policies $(\chi^\alpha, \lambda^\alpha)$. Since $\chi^{\alpha(n)}$ and $\lambda^{\alpha(n)}$ are bang-bang policies that, for all $\alpha(n)$ large enough, use $(\overline{\chi}, \overline{\lambda})$ for all large i, we can consider an infinite subsequence $\{n_i\}$ for which $\chi^{\alpha(n_i)} = \chi^1$ and $\lambda^{\alpha(n_i)} = \lambda^1$.

We show that (χ^1, λ^1) are optimal, by showing that no player can benefit by deviating. Assume that player 1 uses χ^1. We consider an auxiliary stochastic game that differs from the original one only in that for all states $i \geq \bar{\imath}$ (given in theorem 4.1), only action $\overline{\chi}$ is available for player 1. It follows that $\{\chi^{\alpha(n)}, \lambda^{\alpha(n)}\}$ are optimal for the new stochastic game as well, for the discount factor $\alpha(n)$, for all n large enough. Since χ^1 is the limit of policies satisfying the structure in theorem 4.1, it follows that $\chi_i^1 = \overline{\chi}$ for all $i \geq \bar{\imath}$. For the new game, it follows from the weak stability that assumptions 1^* and 2(1) hold, see [23], with $\mu_i = x^i$, and hence, by theorem 3.5, (λ^1, χ^1) are optimal for the average cost in the new game. We conclude that, player 2 cannot benefit by deviating from λ^1, if player 1 plays χ^1. This implies that λ^1 is optimal for the average cost in the original game.

Now assume that player 2 uses λ^1. Then player 1 is faced with a negative dynamic programming. Since in this MDP, χ^1 is optimal for the discounted problem for all discount factors $\alpha(n_i)$, $i = 1, 2, \ldots$, it follows from known results (see e.g. [8] theorems 6.1, 6.2 and 7.1) that χ^1 is expected average optimal. This concludes the proof. **QED**

It is simple to relax the conditions on the rate of increase of the function h. We will end this section by briefly describing the main idea of this extension.

First we can relax the condition on h in assumption $4(\alpha)$ as follows. Let $(\overline{\chi}, \overline{\lambda})$ be the policy that chooses actions $\overline{\chi}$ and $\overline{\lambda}$ in all states.

Assumption $4'(\alpha)$: h is convex non-decreasing, such that $V_0^\alpha(\overline{\chi}, \overline{\lambda}) < \infty$.

We can derive the result of theorem 4.1 by the following arguments. By h^N denote the function given by

$$\begin{aligned} h^N(i) &= h(i), & i \le N \\ h^N(i) &= h(N) + (i-N)\big(h(N) - h(N-1)\big), & i > N. \end{aligned}$$

Clearly $h^N \uparrow h$, as N tends to infinity and h^N increases linearly outside a finite set. As a consequence h^N satisfies assumption $4(\alpha)$ for all α.

By virtue of lemma 4.4 in fact the result of theorem 4.1 holds uniformly in N, i.e. there are $\overline{\imath}$ and α^0, such that the assertions of the theorem hold for all values N. By a limiting argument the result holds for h as well, provided that assumption $4'(\alpha)$ holds. Here assumption $4'(\alpha)$ is mainly intended to guarantee finiteness of the limiting values $\lim_{t\to\infty} w_t^\alpha$.

Going back to the average expected cost case, we can relax assumption 5 as follows.

Assumption $5'$:

i) *Weak stability*

$$\overline{\lambda} < \overline{\chi}.$$

ii) h is convex, strictly increasing, such that for some x, $1 < x < \overline{\chi}/\overline{\lambda}$,

$$\sup_{i \in \mathsf{E}} h(i) x^{-i} < \infty.$$

It is easily seen that assumption $5'$ implies assumption $4'(\alpha)$ for all α close to 1, and so theorem 4.5 is valid under the weaker assumption $5'$. Obviously, we could also have taken first the limit $\alpha \to 1$ for the α-discounted problem with the functions h^N, and then the limit $N \to \infty$.

Finally note that the conditions on the rate of increase of h are not essential, as long as the convexity is preserved. This can be proved by similar arguments as we indicated above.

5. References.

[1] Altman, E., Monotonicity of optimal policies in a zero sum game: a flow control model. *Advances of dynamic games and applications*, **1**, 269-286, 1994.

[2] Altman, E., Contrained Markov decision processes. INRIA report No. 2574, 1995.

[3] Altman, E., Flow control using the theory of zero-sum Markov games. *IEEE Trans. Autom. Control*, **39**, No. 4, 814-818, 1994.

[4] Altman, E., Feinberg E. A. and Shwartz A., Weighted discounted zero-sum stochastic games with perfect information, *Submitted for publication.*

[5] Altman and Hordijk, A., Zero-sum Markov games and worst case optimal control of queueing systems. *Queueing Systems* **21**, 415-447, 1995.

[6] Bewley, T. and Kohlberg, E., The asymptotic Theory of Stochastic Games. *Math of Oper. Res.*, **1** No. 3, 1976.

[7] Borkar, V.S. and Ghosh, M.K. Denumerable state stochastic games with limiting average payoff. *JOTA*, 539–560, 1993.

[8] Cavazos-Cadena, R., Recent results on conditions for the existence of average optimal stationary policies. *Ann. Operat. Res.* **28**, 3-28, 1991.

[9] Chung, K.L. *Markov Chains with stationary transition probabilities* (2d edition). Springer Verlag, Berlin, 1967.

[10] Dekker, R. and Hordijk, A. Average, sensitive and Blackwell optimality in denumerable state Markov decision chains with unbounded rewards. *Math. Operat. Res.* **13**, 395-421, 1988.

[11] Dekker, R. and Hordijk, A. Recurrence conditions for average and Blackwell optimality in denumerable state Markov decision chains. *Math. Operat. Res.* **17**, 271-290, 1992.

[12] Dekker, R., Hordijk, A. and Spieksma, F.M. On the relation between recurrence and ergodicity properties in denumerable Markov decision chains. Technical report, Leiden University, Technical report no. TW-91-02, Dept. of Math. & Comp. Sci., Univ. of Leiden (extended version), *Math. Operat. Res.* **19**, No. 3, 539-559, 1994 (shortened version).

[13] Derman, C. and Strauch, R. A note on memoryless rules for controlling sequential control processes. *Ann. Math. Stat.* **37**, 276–278, 1966.

[14] Duyn Schouten, F.A. v.d. *Markov decision drift Processes with continuous time parameter*, Ph.D. thesis, Leiden University, 1979.

[15] Federgruen, A., On N-person stochastic games with denumerable state space. *Adv. Appl. Prob.* **10**, 452-471, 1978.

[16] Federgruen, A., On the functional equations in undiscounted and sensitive discounted stochastic games. *Zeitschrift Operat. Res* **24**, 243-262, 1980.

[17] Federgruen, A., Successive approximation methods in undiscounted stochastic games. *Operat. Res.* **28**, 794-810, 1980.

[18] Glicksberg, I.L. A further generalisation of the Kakutani fixed point theorem, with application to Nash equilibrium points. *Proc. Amer. Math. Soc.* **3**, 170-174, 1952.

[19] Hordijk, A. *Dynamic Programming and Markov Potential Theory.* Mathematical Centre Tract **51**, C.W.I., Amsterdam, 1974.

[20] Hordijk, A., Schweitzer, P.J. and Tijms, H.C. The asymptotic behaviour of the minimal total expected cost for the denumerable Markov decision model. *J. Appl. Prob.* **12**, 298-305, 1975.

[21] Hordijk, A, Vrieze, O.J. and Wanrooij, G.L. Semi-Markov strategies in stochastic games. *Intl. J. Game Th.* **12**, 81-89, 1983.

[22] Hordijk, A. and Duyn Schouten, F.A. v.d. Average optimal policies in Markov decision processes with applications to a queueing and a replacement model. *Adv. Appl. Prob.* **15**, 274-303, 1983.

[23] Hordijk, A. and Spieksma, F.M. On ergodicity and recurrence properties of a Markov chain with an application to an open Jackson network. *Adv. Appl. Prob.*, **24**, 343-376, 1992.

[24] Lippman, S. A., Applying a new device in the optimization of exponential queueing systems. *Opns. Res.* **23**, 687–710, 1975.

[25] Maitra, A. and Sudderth, W., Borel stochastic games with limsup payoff. *The Annals of Probability* **21**, No. 2, 861–885, 1993.

[26] Majumdar, M. and Sundaran, R., Symmetric stochastic games of resource extraction: the existence of non-randomized stationary equilibrium. *Stochastic Games and Related Topics*, 175–190, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.

[27] Mertens, J. -F. and Neyman, A., Stochastic Games. *Internat. J. of Game Theory* **10**, No. 2, 53–66, 1981.

[28] Mertens, J. -F. and Parthasarathy, T., Equilibria for discounted stochastic games. *CORE Discussion Paper No. 8750, Université Catholique de Louvain*, 1987.

[29] Nowak, A. S., Stationary equilibria for nonzero-sum average payoff ergodic stochastic games with general state space. *Advances of dynamic games and applications*, 1994.

[30] Nowak, A. S. and Raghavan, T. E. S., Existence of stationary correlated equilibria with Symmetric information for discounted stochastic games. *Math. Operat. Res.* **17**, 519–526, 1992.

[31] Parthasarathy, T. and Stern, M., Markov games – a survey. Differential Games and Control Theory II, Roxin, Liu and Sternberg, 1977.

[32] Parthasarathy, T. and Sinha, S., Existence of stationary equilibrium strategies in nonzero-sum discounted games with uncountable sate space and state independent transitions, *Internat. J. of Game Theory*, **18**, 189–194, 1989.

[33] Raghavan, T.E.S. and Filar, J.A. Algorithms for stochastic games- A survey. *ZOR*eitschr. Operat. Res. **35**, 437-472, 1991.

[34] Rieder, U., Equilibrium plans for nonzero-sum Markov games. In: *Game Theory and Related Topics*, 91–102, Moeschlin and Pallaschke (eds.) North-Holland, Amsterdam, 1979.

[35] Royden, H.L. *Real Analysis* (3d edition). Macmillan Publishing Company, New York, 1988.

[36] Shapley, L. S., Stochastic games. *Proceeding of the National Academy of Sciences USA* **39**, 1095–1100, 1953.

[37] Schäl, M. On the second optimality equation for semi-Markov decision models. *Math. Operat. Res.*, **17**, 470-486, 1992.

[38] Schweitzer, P.J. Iterative solution to the functional equations of undiscounted Markov renewal programming. *J. Math. Anal. Appl.* **34**, 495-501, 1971.

[39] Sennott, L.I. Zero-sum stochastic games with unbounded costs: discounted and average cost cases. *ZOR* **40**, 145–162, 1994.

[40] Serfozo, R. F. An equivalence between continuous and discrete time Markov Decision Processes. *Opns. Res.* **27**, No. 3, 616–620, 1978.

[41] Spieksma, F.M. *Geometrically ergodic Markov Chains and the optimal Control of Queues*, Ph. D. thesis, Leiden University, 1990 (available on request from the author).

[42] Spieksma, F.M. and Tweedie, R.L. Strengthening ergodicity to geometric ergodicity of Markov chains. *Stoch. Mod.* **10**, 45-75, 1994.

[43] Vrieze, O. J. Stochastic games with finite state and action spaces. Math. Centrum, Amsterdam, 1983.

[44] Wal, J. v.d., *Stochastic Dynamic Programming: successive approximations and nearly optimal strategies for Markov decision processes and Markov games.* Ph.D. thesis, Eindhoven, 1980.

[45] Wal, J. v.d., Successive Approximations for Average Reward Markov Games, *Internar. J. of Game Theory*, **9**, No. 1, 13–24.

[46] Wessels, J. Markov games with unbounded rewards. Memorandum COSOR 76-05; in: M. Schäl (edt) *Dynamische Optimierung*, Bonner Math. Schriften **98**, Bonn, 1977.