WEIGHTED DISCOUNTED STOCHASTIC GAMES WITH PERFECT INFORMATION

Eitan ALTMAN
INRIA, B.P.93
2004 Route des Lucioles
06902 Sophia-Antipolis Cedex, France

Eugene A. FEINBERG
Harriman School for Management and Policy
SUNY at Stony Brook
Stony Brook, NY 11794-3775, U.S.A.

Adam SHWARTZ
Department of Electrical Engineering
Technion—Israel Institute of Technology
Haifa 32000, Israel

February 1997
Revised: February 1998

Abstract

We consider a two-person zero-sum stochastic game with an infinite time horizon. The payoff is a linear combination of expected total discounted rewards with different discount factors. For a model with a countable state space and compact action sets, we characterize the set of persistently optimal (sub-game perfect) policies. For a model with finite state and action sets and with perfect information, we prove the existence of an optimal pure Markov policy, which is stationary from some epoch onward, and we describe an algorithm to compute such a policy. We provide an example which shows that an optimal policy, which is stationary after some step, may not exist for weighted discounted sequential games with finite state and action sets and without the perfect information assumption. We also provide examples of similar phenomena of nonstationary behavior for the following two classes of problems with weighted discounted criteria: (i) models with one controller and with finite state and compact action sets, (ii) nonzero-sum games with perfect information and with finite state and action sets.

1 Introduction

Several problems in finance, project management, budget allocation, production, and management of computer and telecommunication systems lead to sequential decision problems where the objective functions are linear combinations of total expected discounted rewards, each with its own discount factor. Some of these problems are described in [4, 5, 6, 15, 24].

Many applications of these criteria arise from the conflict of short- and long-term interests, since discount factors may be viewed as related to time-scales. For example, in the development or selection of technologies, ecological consequences and objectives have significantly larger time-scales than direct economical goals. This is also the case in applications to pavement management in roads [12] and building maintenance [28]. Other applications are related to control of systems in which different components have different reliability. Often, different system or performance measures such as costs of items change over time in different ways (see [6]). Production costs often change due to learning or experience (see [5]), at a rate which is unrelated to the change in cost of different parts, etc.

Markov Decision Processes with weighted criteria have been studied in [3, 4, 5, 6, 8, 15]. Even in this case, when there is just one player, the results for problems with weighted discounted rewards differ significantly from the results for standard discounted models. For example, stationary optimal policies may not exist for weighted discounted problems with finite state and action sets [4]. However, in the case of one player, there exist optimal pure Markov policies which are stationary from some epoch onward [4].

Stochastic two-person zero-sum weighted discounted games have been studied in [9] where the existence of ϵ-optimal policies which are stationary from some epoch onward is proved. The existence of a value was established in [4, 9] by reducing weighted discounted games to standard discounted games with expanded state spaces. The existence of the value can also be proved by using the results of [21], which studies non-stationary stochastic games.

This paper deals with a two-person zero-sum stochastic game with weighted discounted payoffs. The main goal is to study finite state and action models with perfect information.

Perfect information means that at any state either the action set of player 1 or the action set of player 2 is a singleton (see [11]). We show that for each player there exists an optimal pure Markov policy which is stationary from some epoch onward. We also provide an algorithm that computes such a policy in a finite number of steps. An optimal policy which is stationary after some step may not exist for standard weighted discounted games with finite state and action sets.

The paper is organized as follows. The model definition and the notation are given in Section 2. Section 3 introduces and studies lexicographical games. In particular, in Section 3 we describe the structure of lexicographically optimal policies for discounted stochastic games with countable state and compact action sets. In section 4 we describe general results on countable state weighted discounted games. Section 5 deals with finite state and action stochastic games with perfect information. We prove the existence of an optimal pure Markov policy which is stationary from some step onward, describe the sets of all optimal policies, and formulate an algorithm for their computation. Section 6 deals with counterexamples.

As counterexamples in [4] and in Section 6 show, the main results of the paper, presented in Section 5, hold just for games with perfect information, with finite state and action sets, and a zero-sum assumption on the costs. The results on lexicographical games and on the existence of optimal Markov policies, presented in Sections 3 and 4, hold for countable state games with compact action sets introduced in Section 2.

2 Definitions and Notation

Consider a two-person zero-sum stochastic game with a finite or countable state space $\mathbf{X}$, and two metric spaces of actions $\mathbf{A}$ and $\mathbf{B}$ for players 1 and 2 respectively. The transition probability $p(y|x, a, b)$ is the probability to go to state y given state x, given that actions a and b are chosen by the players. Let $\mathbf{A}(x)$ and $\mathbf{B}(x)$ denote the set of actions available for players 1 and 2 at state x. We assume that $\sum_{y\in\mathbf{X}} p(y|x, a, b) = 1$ for all $x \in \mathbf{X}$, $a \in \mathbf{A}(x)$, and $b \in \mathbf{B}(x)$. In contrast with standard discounted stochastic games (with a single payoff function that player 2 pays player 1 and one discount factor), here we have K payoff functions $r_k : \mathbf{X} \times \mathbf{A} \times \mathbf{B} \to \mathbb{R}$, $k = 1, \ldots, K$, and K discount factors $\beta_k \in [0, 1[$, $k = 1, \ldots, K$, where K is a given positive integer. Throughout the paper the following is in force.

Technical Assumptions. For each $x \in \mathbf{X}$ the sets $\mathbf{A}(x)$ and $\mathbf{B}(x)$ are compact. For each $x, y \in \mathbf{X}$, the transition probabilities $p(y|x, a, b)$ are continuous in a and b, and each function $r_k(x, a, b)$, $k = 1, \ldots, K$ is bounded, upper semi-continuous in $a \in \mathbf{A}(x)$, and lower semi-continuous in $b \in \mathbf{B}(x)$.

Without loss of generality, assume $\beta_1 > \beta_2 > \ldots > \beta_K$. (If the discount factors are not ordered, we reorder them. If $\beta_i = \beta_j$ for some $i > j$, we rename $r_j := r_j + r_i$, eliminate the ith component, and obtain a model with $K - 1$ payoff functions and $K - 1$ discount factors). Define histories $h_t = x_0, a_0, b_0, x_1, a_1, b_1, \ldots, x_t$, where $t = 0, 1, \ldots$. Let $\mathcal{U}$ and $\mathcal{V}$ be the set of (behavioral) policies available to players 1 and 2 respectively. Policies $u \in \mathcal{U}$ and $v \in \mathcal{V}$ are sequences $u = u_0, u_1, \ldots$ and $v = v_0, v_1, \ldots$, where u_t and v_t are probability distributions respectively on $\mathbf{A}(x_t)$ and $\mathbf{B}(x_t)$ conditioned on h_t. The randomizations used by the two players are assumed to be independent.

A policy u for player 1 is (randomized) Markov if at any time t, u_t depends only on the current state x_t. A Markov policy is called stationary if it is time homogeneous, i.e. $u_0 = u_1 = \ldots$. A policy u for player 1 is called pure if the distribution $u_t(\cdot|h_t)$ is concentrated at one point $u_t(h_t)$ for each history h_t, $t = 0, 1, \ldots$. Pure stationary policies are called deterministic. For $N = 0, 1, \ldots$, a pure Markov policy u for player 1 is called (N, ∞)-stationary if $u_t = u_N$ for all $t \geq N$. The notions of $(0, \infty)$-stationary and deterministic policies coincide. Special classes of policies for player 2 are defined analogously.

Given an initial state x, each pair of policies (u, v) defines a probability measure $P_x^{u,v}$ on the set of trajectories $x_0, a_0, b_0, x_1, a_1, b_1, \ldots$. We denote by $E_x^{u,v}$ the expectation with respect to this measure.

The discounted payoff associated with the one-step payoff r_k and discount factor β_k for an initial state x, where the players use policies u and v, is defined to be

$$V_k(x, u, v) = E_x^{u,v} \sum_{t=0}^{\infty} (\beta_k)^t r_k(x_t, a_t, b_t). \tag{1}$$

The weighted discounted payoff corresponding to the initial state x, and policies u and v is then given by

$$V(x, u, v) = \sum_{k=1}^{K} V_k(x, u, v). \tag{2}$$

Player 1 wishes to maximize $V(x, u, v)$, and player 2 wishes to minimize it.

Remark 2.8 in [4] reduces this game to a game with one discount factor and with a countable state space. Countable state discounted games with compact action sets have values; see e.g. [7], [19], or Theorem 3.1. Therefore, countable state games with compact action sets and with weighted discounted payoffs have values as well. A reduction to a game with one discount factor but with a continuous state space was described in [9]. Define $V(x)$ to be the value of the weighted discounted game and, for $k = 1, \ldots, K$, let $V_k(x)$ denote the value of the game with criterion $V_k(x, \cdot, \cdot)$.

A policy u^* is said to be optimal for player 1 in game (2) if and only if for any $u \in \mathcal{U}$, $\inf_v V(x, u, v) \leq \inf_v V(x, u^*, v)$ (where the latter is equal to $V(x)$) for all $x \in \mathbf{X}$. Optimality of a policy for the second player is defined similarly.

For a policy $u \in \mathcal{U}$ and a history $\tilde{h}_n = \tilde{x}_0, \tilde{a}_0, \tilde{b}_0, \ldots, \tilde{x}_n, \tilde{a}_n, \tilde{b}_n \in (\mathbf{X} \times \mathbf{A} \times \mathbf{B})^{n+1}, n = 0, 1, \ldots$, we define the shifted policy $\tilde{h}_n u$ as the policy which uses, in response to a history $h_m = x_0, a_0, b_0, \ldots, x_m$, the action that u would use at epoch $(n + m)$ if the history $\tilde{h}_n h_m = \tilde{x}_0, \tilde{a}_0, \tilde{b}_0, \ldots, \tilde{x}_n, \tilde{a}_n, \tilde{b}_n, x_0, a_0, \ldots, x_m$ is observed. A similar definition holds for $v \in \mathcal{V}$. We will also use formal the notation $\tilde{h}_{-1}u = u$ and $\tilde{h}_{-1}v = v$. For a Markov policy u of any player $\tilde{h}_n u = (u_{n+1}, u_{n+2}, \ldots)$ does not depend on $\tilde{h}_n$. If u is stationary, $\tilde{h}_n u = u$.

The total expected weighted discounted reward incurred from epoch $(n+1)$, $n = 0, 1, \ldots$, onward if the players use policies u, v, a history $\tilde{h}_n$ took place and $x_{n+1} = x$ is

$$V(x, n+1, u, v) = (\beta_1)^{-(n+1)} \sum_{k=1}^{K} (\beta_k)^{n+1} V_k(x, \tilde{u}, \tilde{v}), \tag{3}$$

where $\tilde{u} = \tilde{h}_n u$, $\tilde{v} = \tilde{h}_n v$. We also set $V(x, 0, u, v) = V(x, u, v)$. We introduce the normalization constant $(\beta_1)^{-n}$ in (3) just in order to have $V_1(x) = V(x, n) + o(1)$.

So any history $\tilde{h}_n$ defines a new stochastic game that starts at epoch $(n+1)$. Let $V(x, n)$ be the value of the zero-sum game that starts at epoch $n = 0, 1, \ldots$ with the payoffs (3). The existence of this value follows from Remark 2.8 in [4]. Since both players know the history and therefore have the same information about the past, this value does not depend on the history before epoch n. A policy u^* (v^*) for player 1 (2) is called persistently optimal [4] if it is optimal and for any $\tilde{h}_n \in (\mathbf{X} \times \mathbf{A})^{n+1}, n = 0, 1, 2, \ldots$, the policy $\tilde{h}_n u$ $(\tilde{h}_n v)$ is optimal (with respect to the cost (3)) as well. We will apply the definition of persistent optimal policies to both criteria V_k, $k = 1, \ldots, K$, and V.

The main objective of this paper is to study games with perfect information which are a special case of stochastic games (see e.g. [11, 16]). We say that a game is with perfect information if there exist two sets of states $\mathbf{Y}$ and $\mathbf{Z}$ such that: (i) $\mathbf{Y} \cup \mathbf{Z} = \mathbf{X}$, (ii) $\mathbf{Y} \cap \mathbf{Z} = \emptyset$, and (iii) the sets $\mathbf{A}(z)$ and $\mathbf{B}(y)$ are singletons for all $z \in \mathbf{Z}$ and for all $y \in \mathbf{Y}$. In particular, if $p(\mathbf{Y}|y, a, b) = p(\mathbf{Z}|z, a, b) = 0$ for all $y \in \mathbf{Y}$, $z \in \mathbf{Z}$, $a \in \mathbf{A}$, and $b \in \mathbf{B}$ in a game with perfect information, then the players make their moves sequentially.

A particular, important example is a stochastic game where the players make their moves simultaneously, but player 2 knows the decision of player 1 at each epoch [16]. In other words, v_t may depend on (h_t, a_t), not just on h_t. In this game, all definitions of special policies should be modified by replacing x_t with (x_t, a_t) in all conditional distributions v_t. Let us define an equivalent game with perfect information. All objects in this new model are marked with "~." Let $\tilde{\mathbf{Y}} = \mathbf{X}$, $\tilde{\mathbf{Z}} = \mathbf{X} \times \mathbf{A}$, $\tilde{\mathbf{X}} = \tilde{\mathbf{Y}} \cup \tilde{\mathbf{Z}}$, and $\tilde{\mathbf{A}}(x) = \mathbf{A}(x)$, $\tilde{\mathbf{B}}(x, a) = \mathbf{B}(x)$, $\tilde{\mathbf{A}}(x, a) = \{a\}$, and $\tilde{\mathbf{B}}(x)$ be any singleton, where $x \in \mathbf{X}$ and $a \in \mathbf{A}$. We define transition probabilities $\tilde{p}$ and payoff functions $\tilde{r}_k$

which do not depend on the action of player 1 (2) on $\mathbf{Z}$ ($\mathbf{Y}$). For $a \in \mathbf{A}$ and $b \in \mathbf{B}$ we set

$$\tilde{p}(\tilde{y}|\tilde{x}, a, b) = \begin{cases} 1 & \text{if } \tilde{x} = x \in \mathbf{X},\ \tilde{y} = (x, a), \\ p(y|x, a, b) & \text{if } \tilde{x} = (x, a) \in \mathbf{X} \times \mathbf{A},\ \tilde{y} = y \in \mathbf{X}, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\tilde{r}_k(\tilde{x}, a, b) = \begin{cases} r_k(x, a, b) & \text{if } \tilde{x} = (x, a) \in \mathbf{X} \times \mathbf{A}, \\ 0 & \text{otherwise.} \end{cases}$$

Each step in the original model corresponds to two sequential steps in the new one. In order to get the same total payoffs for initial states from $\mathbf{X}$, we set $\tilde{\beta}_k = \sqrt{\beta_k}$. It is easy to see that, for all initial points from $\mathbf{X}$, there is a one-to-one correspondence between policies in these two models and $\tilde{V}_k(x, u, v) = V_k(x, u, v)$ and therefore $\tilde{V}(x, u, v) = V(x, u, v)$ for all policies u, v and for all states $x \in \mathbf{X}$.

3 Lexicographical Stochastic Discounted Games

Denote by $\mathrm{B}(E)$ the Borel σ-field on a metric space E, and by $\mathcal{P}(E)$ the set of probability distributions on $(E, \mathrm{B}(E))$, endowed with the weak topology. If E is compact then $\mathcal{P}(E)$ is compact in this topology [22].

Fix x and let $f(a, b)$ be a bounded function on $\mathbf{A} \times \mathbf{B}$ which is upper semi-continuous in a on $\mathbf{A}(x)$ and lower semi-continuous in b on $\mathbf{B}(x)$. Let $b(a)$ and f satisfy $f(a) = \min_b f(a, b) = f(a, b(a))$. Since $a_n \to a$ implies $f(a) = f(a, b(a)) \geq \limsup_{n\to\infty} f(a_n, b(a)) \geq \limsup_{n\to\infty} f(a_n, b(a_n)) = \limsup_{n\to\infty} f(a_n)$, we conclude that $f(a)$ is upper semi-continuous. Similarly $f(a) = \max_a f(a, b)$ is lower semi-continuous. The function $f(p, q) = \int_{\mathbf{A}(x)} \int_{\mathbf{B}(x)} f(a, b) p(da) q(db)$ is upper semi-continuous on $\mathcal{P}(\mathbf{A}(x))$ and it is lower semi-continuous on $\mathcal{P}(\mathbf{B}(x))$ [2, p. 17], convex in p and concave in q (actually, it is linear in each of these coordinates).

For each $x \in \mathbf{X}$ let $\mathcal{A}(x)$ in $\mathcal{P}(\mathbf{A}(x))$ and $\mathcal{B}(x)$ in $\mathcal{P}(\mathbf{B}(x))$ be nonempty convex compact subsets. In particular, one may consider $\mathcal{A}(x) = \mathcal{P}(\mathbf{A}(x))$ and $\mathcal{B}(x) = \mathcal{P}(\mathbf{B}(x))$. We denote by $\mathcal{U}_{\mathcal{A}}$ ($\mathcal{V}_{\mathcal{B}}$) the set of policies for player 1 (2) such that $u_t(\cdot|h_t) \in \mathcal{A}(x_t)$ ($v_t(\cdot|h_t) \in \mathcal{B}(x_t)$) for all $h_t = x_0, a_0, b_0, \ldots, x_t$, $t = 0, 1, \ldots$. We notice that $\mathcal{U} = \mathcal{U}_{\mathcal{A}}$ if and only if $\mathcal{A}(x) = \mathcal{P}(\mathbf{A}(x))$ for all $x \in \mathbf{X}$. Similarly, $\mathcal{V} = \mathcal{V}_{\mathcal{B}}$ if and only if $\mathcal{B}(x) = \mathcal{P}(\mathbf{B}(x))$ for all $x \in \mathbf{X}$.

By Theorem 3.4 in Sion [27]

$$\max_{p \in \mathcal{A}(x)} \min_{q \in \mathcal{B}(x)} f(x, p, q) = \min_{q \in \mathcal{B}(x)} \max_{p \in \mathcal{A}(x)} f(x, p, q) \tag{4}$$

and the appropriate minimums and maximums exist in (4). We denote

$$\mathbf{val}\ f(x,\mathcal{A},\mathcal{B}) = \min_{q\in\mathcal{B}(x)} \max_{p\in\mathcal{A}(x)} f(x,p,q). \tag{5}$$

Let

$$\mathbf{P}(x,f,\mathcal{A},\mathcal{B}) \quad = \{p\in\mathcal{A}(x) : \min\nolimits_{q\in\mathcal{B}(x)} f(x,p,q) = \mathbf{val}\ f(x)\}, \tag{6}$$

$$\mathbf{Q}(x,f,\mathcal{A},\mathcal{B}) \quad = \{q\in\mathcal{B}(x) : \max\nolimits_{p\in\mathcal{A}(x)} f(x,p,q) = \mathbf{val}\ f(x)\}. \tag{7}$$

When $\mathcal{A}(y) = \mathcal{P}(\mathbf{A}(y))$ and $\mathcal{B}(y) = \mathcal{P}(\mathbf{B}(y))$ for all $y\in\mathbf{X}$ we use the notation $\mathbf{val}\ f(x) = \mathbf{val}\ f(x,\mathcal{A},\mathcal{B})$, $\mathbf{P}(x,f) = \mathbf{P}(x,f,\mathcal{A},\mathcal{B})$, and $\mathbf{Q}(x,f) = \mathbf{Q}(x,f,\mathcal{A},\mathcal{B})$.

Since $\min_q f(x,p,q)$ is upper semi-continuous in p, $\mathbf{P}(x,f,\mathcal{A},\mathcal{B})$ are nonempty and compact. In addition, they are convex; see Lemma 2.1.1 Karlin [14] for the proof of a similar statement. Similarly, $\mathbf{Q}(x,f,\mathcal{A},\mathcal{B})$ are nonempty convex compact sets. We have that $p\in\mathcal{A}(x)$ ($q\in\mathcal{B}(x)$) is an optimal policy for player 1 (2) in a zero-sum game with the payoff function $f(x,a,b)$ and sets of decisions limited to randomized decisions from $\mathcal{A}(x)$ and $\mathcal{B}(x)$ if and only if $p\in\mathbf{P}(x,f,\mathcal{A},\mathcal{B})$ ($q\in\mathbf{Q}(x,f,\mathcal{A},\mathcal{B})$). If $\mathbf{A}(x)$ and $\mathbf{B}(x)$ are finite then $\mathbf{P}(x,f)$ and $\mathbf{Q}(x,f)$ are nonempty polytopes. This fact is known as the Shapley-Snow theorem [23], [26].

For stochastic games with one discount factor ($K=1$), we omit the subscripts $k=1$ from the notation. In that case, the game reduces to a standard stationary (i.e. with a time-homogeneous structure) stochastic game, for which a very rich theory exists see e.g. [10, 13, 18, 19, 20, 23, 25]. We shall make use of the following known theorem.

Theorem 3.1 *Let $K=1$. Consider a zero-sum discounted stochastic game such that the set of policies for player 1 (2) is $\mathcal{U}_{\mathcal{A}}$ ($\mathcal{V}_{\mathcal{B}}$), where $\mathcal{A}(x)$ ($\mathcal{B}(x)$) are nonempty convex compact subsets of $\mathcal{P}(\mathbf{A}(x))$ ($\mathcal{P}(\mathbf{B}(x))$) defined for all $x\in\mathbf{X}$.*

(i) The game has a value $V(x) = V(x,\mathcal{A},\mathcal{B})$ which is the unique bounded solution of

$$V(x,\mathcal{A},\mathcal{B}) = \mathbf{val}\ F(x,\mathcal{A},\mathcal{B}), \qquad x\in\mathbf{X}, \tag{8}$$

with

$$F(x,a,b) = r(x,a,b) + \beta\sum_{z\in\mathbf{X}} p(z|x,a,b)V(z). \tag{9}$$

(ii) A stationary policy u^* (v^*) for player 1 (2) is optimal if and only if $u^*(\cdot|x)\in\mathbf{P}(x,F,\mathcal{A},\mathcal{B})$ ($v^*(\cdot|x)\in\mathbf{Q}(x,F,\mathcal{A},\mathcal{B})$) for all $x\in\mathbf{X}$.

(iii) A policy u^* (v^*) for player 1 (2) is persistently optimal if and only if $u_n^*(\cdot|h_n) \in \mathbf{P}(x_n, F, \mathcal{A}, \mathcal{B})$ ($v_n^*(\cdot|h_n) \in \mathbf{Q}(x_n, F, \mathcal{A}, \mathcal{B})$) for all $h_n = x_0, a_0, b_0, \ldots, x_n$, $n = 0, 1, \ldots$.

Proof. (i, ii) For the case of a standard game when $\mathcal{A}(x) = \mathcal{P}(\mathbf{A}(x))$ and $\mathcal{B}(x) = \mathcal{P}(\mathbf{B}(x))$ for all $x \in \mathbf{X}$, these statements are well-known. Indeed, (i) and the first part of (ii) are a particular case of a corresponding results for games with Borel state spaces; see e.g. [17, 19]. Note that the immediate costs in [17] are assumed to be continuous; however, the proof extends in a straight forward way to our case, since it is based on minimax results that hold also in our case. Note also that it is stated in Theorem 1 in [17] that V is the unique solution of (8); this is in general not true. However, the proof of Theorem 1 in [17] shows that V is the unique *bounded* solution of (8). The "only if" part of (ii) follows from Theorem 2.3 (i) in [1] (again, continuous immediate costs are considered, but the proof holds also under our assumptions).

If $\mathcal{A}(x) \neq \mathcal{P}(\mathbf{A}(x))$ or $\mathcal{B}(x) \neq \mathcal{P}(\mathbf{B}(x))$ for some $x \in \mathbf{X}$, we consider the game with the action sets $\mathcal{A}(x)$ and $\mathcal{B}(x)$, $x \in \mathbf{X}$. Since the reward function $r(x, p, q)$ is concave in p and convex in q, this game has a solution in pure policies. Therefore, statements (i) and (ii) for the case $\mathcal{A}(x) = \mathcal{P}(\mathbf{A}(x))$ and $\mathcal{B}(x) = \mathcal{P}(\mathbf{B}(x))$ for all $x \in \mathbf{X}$ imply the corresponding statements for nonempty convex compact sets $\mathcal{A}(x)$ and $\mathcal{B}(x)$, $x \in \mathbf{X}$.

(iii) Let u^* and v^* be as stated. Consider a game with a finite horizon T and terminal cost V, which is the unique bounded solution of (8). Consider the value of the game from time n till time T, which we call the (n, T) game, given that a history $\tilde{h}_{n-1}$ took place and $x_n = x$. It easily follows from a backward induction argument that the value of this game is V, and that u^* and v^* are optimal. Since this holds for any T and since the immediate cost is bounded, a standard limiting argument shows that the value of the (n, ∞) game is also V, and that u^* and v^* are optimal for the (n, ∞) game as well. This establishes the "if" part.

To show the "only if" part, consider some history $\tilde{h}_{n-1}$ and some state $x_n = x$, and let u be a policy for player 1 which does not satisfy the condition on u^* for that history and x. Then

$$\begin{aligned} V(x, n, u, v^*) &\leq E^{u, v^*}\left[r(x, A_n, B_n) + \beta \sum_{z \in \mathbf{X}} p(z|x, A_n, B_n) V(z) \right] \\ &< \mathbf{val} F(x, \mathcal{A}, \mathcal{B}) = V. \end{aligned}$$

This establishes the "only if" part for player 1. A symmetric argument leads to the result for player 2. ■

For one-step as well as for stochastic zero-sum games, let us describe the notions of lexicographical games and lexicographical values: the formal definitions are given below. Consider a game with sets $\mathcal{U}_1$ and $\mathcal{V}_1$ of policies for players 1 and 2 and with a vector of payoffs $(\tilde{V}_1(x,u,v),\ldots,\tilde{V}_K(x,u,v))$. We say that a vector $\tilde{V}(x)=(\tilde{V}_1(x),\ldots,\tilde{V}_K(x))$ is a lexicographical value of this game if (i) $\tilde{V}_1(x)$ is the value for the game Γ_1 with sets of policies $\mathcal{U}_1$ and $\mathcal{V}_1$ for players 1 and 2 and with payoff $\tilde{V}_1(x,u,v)$, and (ii) for $k=1,\ldots,K-1$, $\tilde{V}_{k+1}(x)$ is the value of the game Γ_{k+1} whose set of policies consists of those policies which are optimal for the game Γ_k, and whose payoff function is $\tilde{V}_{k+1}(x,u,v)$. A policy is called lexicographically optimal if it is optimal for the game Γ_K.

First, we give definitions for one-step games and construct the sets of lexicographically optimal policies for players 1 and 2. Then we we shall define lexicographically persistently optimal policies for stochastic games.

Consider K payoff functions $f_1,\ldots,f_K$, where $f_k=f_k(x,a,b)$ with $a\in\mathbf{A}$, $b\in\mathbf{B}$. All these functions are assumed to be bounded, upper semi-continuous in a, and lower semi-continuous in b. Given $x\in\mathbf{X}$, we define lexicographically optimal policies for games with these payoffs. The sets of policies for player 1 and 2 in game Γ_1 are respectively $\mathcal{P}(\mathbf{A}(x))$ and $\mathcal{P}(\mathbf{B}(x))$ which are nonempty convex compact sets in the weak topology. Therefore, the sets of optimal policies for this game with payoff f_1 are nonempty convex compact subsets of $\mathcal{P}(\mathbf{A}(x))$ and $\mathcal{P}(\mathbf{B}(x))$. We consider the game Γ_2 with these sets of policies and payoff function f_2. The sets of optimal policies for this game are also nonempty, convex, and compact. By repeating this procedure, we define the vector value of the lexicographical game and the set of optimal policies. Combining lexicographical optimal policies for one-step games with Theorem 3.1, we shall construct the value and the sets of persistently optimal policies for a standard stochastic discounted game.

Now we give formal definitions. We start with a one-step game. Consider an arbitrary $x\in\mathbf{X}$. We denote $\mathbf{val}_1\ f_1(x)=\mathbf{val}\ f(x)$, $\mathbf{P}_1(x,f_1)=\mathbf{P}(x,f_1)$, and $\mathbf{Q}_1(x,f_1)=\mathbf{Q}(x,f_1)$. For fixed $k=1,\ldots,K-1$, suppose that the value $\mathbf{val}_k\ f_k(x)$ and two nonempty convex compact sets $\mathbf{P}_k(x,f_k)\subseteq\mathcal{P}(\mathbf{A}(x))$ and $\mathbf{Q}_k(x,f_k)\subseteq\mathcal{P}(\mathbf{B}(x))$ are given.

Let $\mathcal{A}_k(x)=\mathbf{P}_k(x,f_k)$ and $\mathcal{B}_k(x)=\mathbf{Q}_k(x,f_k)$. We define the value

$$\mathbf{val}_{k+1}\ f_{k+1}(x)=\mathbf{val}\ f_{k+1}(x,\mathcal{A}_k,\mathcal{B}_k)$$

and the sets of optimal policies

$$\mathbf{P}_{k+1}(x,f_{k+1})=\mathbf{P}(x,f_{k+1},\mathcal{A}_k,\mathcal{B}_k),$$

$$\mathbf{Q}_{k+1}(x,f_{k+1})=\mathbf{Q}(x,f_{k+1},\mathcal{A}_k,\mathcal{B}_k),$$

which are nonempty, convex, and compact. This construction implies that $(\mathbf{val}_1\ f_k(x),\ldots,$ $\mathbf{val}_K\ f_K(x))$ is a lexicographical value of the one-step game and the nonempty convex compact sets $\mathbf{P}_K(x, f_K)$ and $\mathbf{Q}_K(x, f_K)$ are the sets of lexicographically optimal policies for players 1 and 2 respectively.

Now we consider a stochastic game with K discount factors β_k, with K one-step payoff functions r_k, $k = 1, \ldots, K$, and with the payoff criterion $(V_1, \ldots, V_K)$ defined by (1). Here we do not need the assumption $\beta_1 > \ldots > \beta_K$. First we consider this stochastic game with the reward function r_1 and discount factor β_1. In view of Theorem 3.1 (i), this game has a unique value function which we denote by $V_1(x)$, $x \in \mathbf{X}$. Let F_1 be defined by (9) with $r = r_1$, $\beta = \beta_1$, and $V = V_1$. Theorem 3.1 (iii) implies that $\mathcal{U}_{\mathcal{A}_1}$ and $\mathcal{V}_{\mathcal{B}_1}$ are the sets of lexicographically optimal policies for players 1 and 2 respectively where $\mathcal{A}_1(x) = \mathbf{P}(x, F_1)$ and $\mathcal{B}_1(x) = \mathbf{Q}(x, F_1)$ for all $x \in \mathbf{X}$. In addition, $\mathcal{A}_1(x)$ and $\mathcal{B}_1(x)$ are nonempty convex compact sets for all $x \in \mathbf{X}$.

For fixed $k = 1, \ldots, K-1$, suppose that the value $V_k(x)$ and nonempty convex compact sets $\mathcal{A}_k(x)$ and $\mathcal{B}_k(x)$ are defined for all $x \in \mathbf{X}$. We consider a stochastic game with a set of policies $\mathcal{U}_{\mathcal{A}_k}$ for player 1, a set of policies $\mathcal{V}_{\mathcal{B}_k}$ for player 2, and with the payoffs $V_{k+1}(x, u, v)$. Theorem 3.1 (i) implies that this game has a unique value function V_{k+1} which is a unique solution of $V_{k+1}(x) = \mathbf{val}\ F_{k+1}(x, \mathcal{A}_k, \mathcal{B}_k)$ for all $x \in \mathbf{X}$, where F_{k+1} is defined by (9) with $r = r_{k+1}$, $\beta = \beta_{k+1}$, and $V = V_{k+1}$. We denote $\mathcal{A}_{k+1}(x) = \mathbf{P}(x, F_{k+1}, \mathcal{A}_k, \mathcal{B}_k)$ and $\mathcal{B}_{k+1}(x) = \mathbf{Q}(x, F_{k+1}, \mathcal{A}_k, \mathcal{B}_k)$. By Theorem 3.1 (iii), $\mathcal{U}_{\mathcal{A}_{k+1}}$ and $\mathcal{V}_{\mathcal{B}_{k+1}}$ are the sets of persistently optimal policies for this game.

We say that a policy u (v) for player 1 (2) is lexicographically persistent optimal if $u \in \mathcal{U}_{\mathcal{A}_K}$ ($v \in \mathcal{V}_{\mathcal{B}_K}$). The above construction and Theorem 3.1 lead to the following theorem in which we do not assume that $\beta_1 > \ldots > \beta_K$.

Theorem 3.2 *Consider a stochastic zero-sum game with K reward functions $r_1, \ldots, r_K$, and with K discount factors $\beta_1, \ldots, \beta_K$.*

(i) This game has a lexicographical value $V_1, \ldots, V_K$, where V_k, $k = 1, \ldots, K$, is a unique solution of $V_k(x) = \mathbf{val}_k\ F_k(x)$ with $F_k(x)$ defined for all $x \in \mathbf{X}$ by (9) with $r = r_k$, $V = V_k$, and with $\beta = \beta_k$, $k = 1, \ldots, K$.

(ii) A stationary policy u^* (v^*) for player 1 (2) is lexicographically optimal if and only if $u^*(\cdot|x) \in \mathbf{P}_K(x, F_K)$ ($v^*(\cdot|x) \in \mathbf{Q}_K(x, F_K)$) for all $x \in \mathbf{X}$.

(iii) A policy u^* (v^*) for player 1 (2) is lexicographically persistently optimal if and only if $u^*_n(\cdot|h_n) \in \mathbf{P}_K(x_n, F_K)$ ($v^*_n(\cdot|h_n) \in \mathbf{Q}_K(x_n, F_K)$) for all $h_n = x_0, a_0, b_0, \ldots, x_n$, $n = 0, 1, \ldots$.

Now we consider a game with perfect information. Without loss of generality, we can consider the situation that the singletons $\mathbf{B}(x)$ $(\mathbf{A}(x))$ coincide for all $x \in \mathbf{Y}$ $(x \in \mathbf{Z})$. If we write a triplet (x, a, b), this means that $\{a\} = \mathbf{A}(x)$ for $x \in \mathbf{Z}$ and $\{b\} = \mathbf{B}(x)$ for $x \in \mathbf{Y}$. For a stochastic discounted zero-sum game with perfect information (8) can be rewritten in the following form

$$V(x) = \mathbf{val}\ F(x) = \begin{cases} \max_{a \in \mathbf{A}(x)} F(x,a,b) & \text{if } x \in \mathbf{Y}, \\ \min_{b \in \mathbf{B}(x)} F(x,a,b) & \text{if } x \in \mathbf{Z}, \end{cases}$$

where F is defined in (9).

Let

$$\mathbf{A}(x,F) = \{a \in \mathbf{A}(x) : F(x,a,b) = \mathbf{val}\ F(x)\}, \qquad x \in \mathbf{Y}, \tag{10}$$

$$\mathbf{B}(x,F) = \{b \in \mathbf{B}(x) : F(x,a,b) = \mathbf{val}\ F(x)\}, \qquad x \in \mathbf{Z}. \tag{11}$$

We observe that if $x \in \mathbf{Y}$ then $\mathbf{Q}(x,F)$ is a measure concentrated at the singleton $\mathbf{B}(x)$ and $\mathbf{P}(x,F) = \mathcal{P}(\mathbf{A}(x,F))$. If $x \in \mathbf{Z}$ then $\mathbf{P}(x,F)$ is a measure concentrated at the singleton $\mathbf{A}(x)$ and $\mathbf{Q}(x,F) = \mathcal{P}(\mathbf{B}(x,F))$.

Since a stochastic discounted game with perfect information is a particular case of a general stochastic discounted game, Theorem 3.2 is applicable to games with perfect information. Since the nonempty convex compact sets of optimal policies at each step for games with perfect information are the sets of all randomized policies on subsets of action sets, optimal pure policies exist for games with perfect information. We get the following theorem from Theorem 3.2.

Theorem 3.3 *Consider a zero-sum discounted stochastic game with perfect information.*

(i) A deterministic policy u^* (v^*) for player 1 (2) is optimal if and only if $u^*(x) \in \mathbf{A}(x,F)$ for all $x \in \mathbf{Y}$ $(v^*(x) \in \mathbf{B}(x,F)$ for all $x \in \mathbf{Z})$.

(ii) A pure policy u^* (v^*) for player 1 (2) is persistently optimal if and only if $u_t^*(h_t) \in \mathbf{A}(x_t,F)$ whenever $x_t \in \mathbf{Y}$ $(v_t^*(h_t) \in \mathbf{B}(x_t,F)$ whenever $x_t \in \mathbf{Z})$ for all $h_t = x_0, a_0, b_0, \ldots, x_t$, $t = 0, 1, \ldots$.

Now we consider a lexicographical zero-sum stochastic discounted game with perfect information. We define the sets $\mathbf{A}_1(x,F_1) = \mathbf{A}(x,F)$ and $\mathbf{B}_1(x,F_1) = \mathbf{B}(x,F)$ with $r = r_1$ and $\beta = \beta_1$. We also set for $k = 1, \ldots, K-1$

$$\mathbf{A}_{k+1}(x, F_{k+1}) = \{a' \in \mathbf{A}_k(x) : F_{k+1}(x, a', b) = \max_{a \in \mathbf{A}_k(x)} F_{k+1}(x, a, b)\}, \qquad x \in \mathbf{Y}, \tag{12}$$

$$\mathbf{B}_{k+1}(x, F_{k+1}) = \{b' \in \mathbf{B}_k(x) : F_{k+1}(x, a, b') = \min_{b \in \mathbf{B}_k(x)} F_{k+1}(x, a, b)\}, \qquad x \in \mathbf{Z}. \tag{13}$$

Then if $x \in \mathbf{Y}$ we have that $\mathbf{Q}_k(x, F_k)$ is a measure concentrated at the singleton $\mathbf{B}(x)$ and $\mathbf{P}_k(x, F_k) = \mathcal{P}(\mathbf{A}_k(x, F_k))$, $k = 1, \ldots, K$. If $x \in \mathbf{Z}$ then $\mathbf{P}_k(x, F_k)$ is a measure concentrated at the singleton $\mathbf{A}(x)$ and $\mathbf{Q}_k(x, F_k) = \mathcal{P}(\mathbf{B}_k(x, F_k))$, $k = 1, \ldots, K$.

For games with perfect information, lexicographically optimal policies can be selected among pure policies. The following corollary follows from Theorem 3.2 and it does not assume that $\beta_1 > \ldots > \beta_K$.

Corollary 3.1 *Consider a stochastic zero-sum game with perfect information, with K reward functions $r_1, \ldots, r_K$, and with K discount factors $\beta_1, \ldots, \beta_K$.*

(i) A deterministic policy u^* (v^*) for player 1 (2) is lexicographically optimal if and only if $u^*(x) \in \mathbf{A}_K(x, F_K)$ for all $x \in \mathbf{Y}$ ($v^*(x) \in \mathbf{B}_K(x, F_K)$ for all $x \in \mathbf{Z}$).

(ii) A pure policy u^* (v^*) for player 1 (2) is lexicographically persistently optimal if and only if $u_t^*(h_t) \in \mathbf{A}_K(x_t, F_K)$ for all $x_t \in \mathbf{Y}$ ($v_t^*(h_t) \in \mathbf{B}_K(x_t, F_K)$ for all $x_t \in \mathbf{Z}$) for all $h_t = x_0, a_0, b_0, \ldots, x_t$, $t = 0, 1, \ldots$.

4 Countable State Weighted Discounted Games

Weighted discounted stochastic games are special case of nonstationary stochastic games, as defined in [21]. One can thus use the results in [21] to establish the existence of a value. More insight into the structure of the game can be gained, however, by reducing the game into a standard equivalent discounted stochastic game, as was established in Feinberg and Shwartz [4], Remark 2.8, and in Filar and Vrieze [9]. They showed that a weighted discounted stochastic game can be reduced to a standard discounted stochastic game with discount factor β_1, state space $\overline{\mathbf{X}} = \mathbf{X} \times \{0, 1, \ldots\}$, action sets $\overline{\mathbf{A}}(x, n) = A(x)$ and $\overline{\mathbf{B}}(x, n) = B(x)$, one-step payoffs

$$\overline{r}(x, n, a, b) = \sum_{k=1}^{K} \left(\frac{\beta_k}{\beta_1}\right)^n r_k(x, a, b), \qquad n = 0, 1, \ldots, \tag{14}$$

and transition probabilities

$$\overline{p}((x,n),a,b,(y,k)) = \begin{cases} p(x,a,b,y) & \text{if } k = n+1, \\ 0 & \text{otherwise.} \end{cases}$$

Actually, the new game is equivalent to the set of original games that starts at all possible epochs $n = 0, 1, \ldots$, not just at $n = 0$. For a one-step game with a payoff function f that depends on a parameter other than x, we also consider a notion of a value. In particular, we consider $f = f(x,n)$. We also can consider the sets of optimal policies $\mathbf{P}(x,n,f)$ and $\mathbf{Q}(x,n,f)$ defined by (6) and (7) when f depends on n. The following theorem follows from Theorem 3.1 above and from Remark 2.8 in [4].

Theorem 4.1 *Consider a weighted discounted zero-sum Markov game.*

(i) Each of the games that start at epochs $n = 0, 1, \ldots$ has a value $V(x,n)$ which is the unique solution of

$$V(x,n) = \mathbf{val}\ F(x,n), \qquad x \in \mathbf{X},\ n = 0, 1, \ldots, \tag{15}$$

with

$$F(x,n,a,b) = \overline{r}(x,n,a,b) + \beta_1 \sum_{z \in \mathbf{X}} p(z|x,a,b) V(z,n+1). \tag{16}$$

(ii) A policy u^* (v^*) for player 1 (2) is persistently optimal if and only if $u_t^*(\cdot|h_t) \in \mathbf{P}(x_t,t,F)$ ($v_t^*(\cdot|h_t) \in \mathbf{Q}(x_t,t,F)$) for all $h_t = x_0, a_0, b_0, \ldots, x_t$, $t = 0, 1, \ldots$.

Since policies u and v may be selected to be Markov in Theorem 4.1 (ii), this theorem implies the following result.

Corollary 4.1 *In a weighted discounted zero-sum stochastic game each player has an optimal Markov policy.*

Now we consider a weighted discounted zero-sum stochastic game with perfect information. We observe that the game with state space $\overline{\mathbf{X}}$ is also a game with perfect information, with $\overline{\mathbf{Y}} = \mathbf{Y} \times \{0, 1, \ldots\}$ and $\overline{\mathbf{Z}} = \mathbf{Z} \times \{0, 1, \ldots\}$. Since F depends on n, the sets of optimal actions defined in (10) and (11) also depend on n. We write $\mathbf{A}(x,n,F)$ and $\mathbf{B}(x,n,F)$. We notice that if $x \in \mathbf{Y}$ then $\mathbf{Q}(x,n,F)$ is a measure concentrated at the singleton $\mathbf{B}(x)$ and $\mathbf{P}(x,n,F) = \mathcal{P}(\mathbf{A}(x,n,F))$. If $x \in \mathbf{Z}$ then $\mathbf{P}(x,n,F)$ is a measure concentrated at the singleton $\mathbf{A}(x)$ and $\mathbf{Q}(x,n,F) = \mathcal{P}(\mathbf{B}(x,n,F))$. Therefore, Theorem 4.1 implies the following result.

Theorem 4.2 *Consider a zero-sum weighted discounted stochastic game with perfect information. A pure policy u^* (v^*) for player 1 (2) is persistently optimal if and only if $u_t^*(h_t) \in \mathbf{A}(x_t, t, F)$ whenever $x_t \in \mathbf{Y}$ ($v_t^*(h_t) \in \mathbf{B}(x_t, t, F)$ whenever $x_t \in \mathbf{Z}$) for all $h_t = x_0, a_0, b_0, \ldots, x_t$, $t = 0, 1, \ldots$.*

Theorem 4.2 implies the following result which is similar to Corollary 4.1.

Corollary 4.2 *In a weighted discounted zero-sum stochastic game with perfect information, each player has an optimal pure Markov policy.*

5 The Finite Case with Perfect Information: Main Results

This section describes the structure of persistently optimal policies in weighted discounted stochastic games with perfect information and with finite state and action sets. By Theorem 4.2 there are sets $\mathbf{A}(x, n, F)$ and $\mathbf{B}(x, n, F)$ such that policies for players 1 and 2 are persistently optimal if and only if at each step they select actions from these sets. The following theorem claims that there exists a finite integer N such that at each state the optimal sets of actions for each player coincide at all steps $n \geq N$. Furthermore, these sets of actions for $n \geq N$ are the sets of lexicographically optimal actions described in Corollary 3.1.

Feinberg and Shwartz [5], Definition 5.4, define a funnel as the set of all policies with the following properties. (Note that there was a typo in that definition: in condition (ii), $\mathbf{A}_n(z)$ should be replaced with $\mathbf{A}_N(z)$). Suppose we are given action sets that depend on the current state and also on time, but for some N, the action sets do not depend on time n, whenever $n \geq N$. The funnel (associated with these action sets) is then the set of all policies that select actions from these action sets.

The following theorem shows that for each player the set of optimal policies is a funnel. Algorithm 5.1 provides a method for computing optimal policies; by Remark 5.1, the algorithm can be used to compute the (time dependent) optimal action sets, and thus obtain the funnel of optimal policies. Recall the assumption $\beta_l > \beta_{l+1}$, $l = 1, \ldots, K-1$.

Theorem 5.1 *Consider a weighted discounted stochastic game with perfect information and with finite state and action sets. There exists a finite integer N such that $A(x, n, F) = A_K(x, F_K)$ and $B(x, n, F) = B_K(x, F_K)$ for all $x \in \mathbf{X}$ and for all $n \geq N$.*

Proof. Let for $k = 1, \ldots, K$

$$C_k = \sup\{|r_k(x, a, b)| : x \in \mathbf{X}, a \in \mathbf{A}(x), b \in \mathbf{B}(x)\}. \tag{17}$$

For $n = 0, 1, \dots$ and for $l = 1, \dots, K-1$, we define

$$\delta_{n,l} = (\beta_l)^{-n} \sum_{k=l+1}^{K} \frac{(\beta_k)^n C_k}{1-\beta_k}. \tag{18}$$

and

$$\gamma_{n,l} = (\beta_l)^{-n} \sum_{k=l+1}^{K} (\beta_k)^n C_k. \tag{19}$$

Observe that $\gamma_{n,l} + \beta_l \delta_{n+1,l} = \delta_{n,l}$ and that $\delta_{n,l} \to 0$ as $n \to \infty$ for all $l = 1, \dots, K-1$.

First we show that there exists $N_{1,1}$ such that $\mathbf{A}(x,t,F) \subseteq \mathbf{A}_1(x,F_1)$ for all $t \geq N_{1,1}$ and for all $x \in \mathbf{X}$. Since all sets $\mathbf{A}(x)$ are singletons for all $x \in \mathbf{Z}$, we have to prove this just for $x \in \mathbf{Y}$.

We observe that for any $x \in \mathbf{X}$, any $n = 0, 1, \dots$, any couple of policies (u, v) for players 1 and 2 respectively, and for any history $\tilde{h}_n = \tilde{x}_0, \tilde{a}_0, \tilde{b}_0, \dots, \tilde{x}_{n-1}, \tilde{a}_{n-1}, \tilde{b}_{n-1}$,

$$|V(x, n, \tilde{h}_{n-1}u, \tilde{h}_{n-1}v) - V_1(x, \tilde{h}_{n-1}u, \tilde{h}_{n-1}v)| \leq \delta_{n,1}. \tag{20}$$

Therefore, $|V(x,n) - V_1(x)| \leq \delta_{n,1}$. From (16), (9), and (14) we have that

$$|F(x,n,a,b) - F_1(x,a,b)| \leq \gamma_{n,1} + \beta_1 \delta_{n+1,1} = \delta_{n,1}. \tag{21}$$

We recall that $\mathbf{B}(x) = \{b\}$ for $x \in \mathbf{Y}$ and $\mathbf{A}(x) = \{a\}$ for $x \in \mathbf{Z}$. Let $x \in \mathbf{Y}$. If $\mathbf{A}(x) = \mathbf{A}_1(x, F_1)$, we set $N_{1,1}(x) = 0$: in particular, $N_{1,1}(x) = 0$ for $x \in \mathbf{Z}$. If $\mathbf{A}(x) \neq \mathbf{A}_1(x, F_1)$, we set

$$N_{1,1}(x) = \min\{n = 0, 1, \dots : \min_{a \in \mathbf{A}(x) \setminus \mathbf{A}_1(x,F_1)} \{V_1(x) - F_1(x,a,b)\} > 2\delta_{n,1}\}. \tag{22}$$

Then for any $a \in \mathbf{A}(x) \setminus \mathbf{A}_1(x, F_1)$ and for any $n \geq N_{1,1}(x)$,

$$\begin{aligned} V(x,n) - F(x,n,a,b) &> V(x,n) - F(x,n,a,b) - V_1(x) + F_1(x,a,b) + 2\delta_{n,1} \geq \\ &(V(x,n) - V_1(x)) + (F_1(x,a,b) - F(x,n,a,b)) + 2\delta_{n,1} \geq 0. \end{aligned} \tag{23}$$

By (23), if $n \geq N_{1,1}(x)$ and $a \notin \mathbf{A}_1(x, F_1)$ then $a \notin \mathbf{A}(x, n, F)$. In other words, $\mathbf{A}_1(x, F_1) \supseteq \mathbf{A}(x, n, F)$ for $n \geq N_{1,1}(x)$.

We repeat the above construction for the second player. For $x \in \mathbf{Z}$ such that $\mathbf{B}_1(x, F_1) \neq \mathbf{B}(x)$, we define

$$N_{2,1}(x) = \min\{n = 0, 1, \dots : \min_{b \in \mathbf{B}(x) \setminus \mathbf{B}_1(x,F_1)} \{F_1(x,a,b) - V_1(x)\} > 2\delta_{n,1}\}. \tag{24}$$

We set $N_{2,1}(x) = 0$ for all other x. Then $\mathbf{B}_1(x, F_1) \supseteq \mathbf{B}(x, n, F)$ for all $n \geq N_{2,1}(x)$, $x \in \mathbf{X}$. We define $N_{i,1} = \max_{x \in \mathbf{X}} N_{i,1}(x)$, $i = 1, 2$.

We set $N_1 = \max\{N_{1,1}, N_{1,2}\}$. Then $\mathbf{A}_1(x, F_1) \supseteq \mathbf{A}(x, n, F)$ and $\mathbf{B}_1(x, F_1) \supseteq \mathbf{B}(x, n, F)$ for all $n \geq N_1$ and for all $x \in \mathbf{X}$. We observe that for $n \geq N_1$ and for any history $\tilde{h}_{n-1}$

$$V_1(x, n, \tilde{h}_{n-1}u, \tilde{h}_{n-1}v) = V_1(x) \tag{25}$$

for any policies u and v that, whenever $t \geq N_1$, select actions from the sets $\mathbf{A}_1(x_t, F_1)$ and $\mathbf{B}_1(x_t, F_1)$.

We consider our game for $n \geq N_1$ and with action sets $\mathbf{A}(\cdot)$ reduced to $\mathbf{A}_1(\cdot, F_1)$ and action sets $\mathbf{B}(\cdot)$ reduced to $\mathbf{B}_1(\cdot, F_1)$. Since in the new model, the component $V_1(x, n, u, v)$ of the payoff function $V(x, n, u, v)$ is constant with respect to policies u and v that start at epoch n, we can remove the first criterion (that is, we can set $r_1 \equiv 0$).

We thereby obtain a model with $(K-1)$ criteria. We repeat this procedure at most $(K-2)$ times and eventually obtain a model with a single payoff function r_K. At each step $l = 2, \ldots, K$, for each $x \in \mathbf{Y}$ such that $\mathbf{A}_{l-1}(x, F_{l-1}) \neq \mathbf{A}_l(x, F_l)$, we define

$$N_{1,l}(x) = \min\{n \geq N_{l-1} : \min_{a \in \mathbf{A}_{l-1}(x, F_{l-1}) \setminus \mathbf{A}_l(x, F_l)} \{F_l(x, a, b) - V_l(x)\} > 2\delta_{n,l}\} \tag{26}$$

and $N_{1,l}(x) = N_{l-1}$ for all other x. For each $x \in \mathbf{Z}$ such that $\mathbf{B}_{l-1}(x, F_{l-1}) \neq \mathbf{B}_l(x, F_l)$, we also define

$$N_{2,l}(x) = \min\{n \geq N_{l-1} : \min_{b \in \mathbf{B}_{l-1}(x, F_{l-1}) \setminus \mathbf{B}_l(x, F_l)} \{V_l(x) - F_l(x, a, b)\} > 2\delta_{n,l}\} \tag{27}$$

and $N_{2,l}(x) = N_{l-1}$ for all other x. We also set $N_{i,l} = \max_{x \in \mathbf{X}} N_{i,l}(x)$, where $i = 1, 2$ and $l = 2, \ldots, K$, and $N_l = \max\{N_{1,l}, N_{2,l}\}$.

After iteration K we have $\mathbf{A}_K(x, F_K) \supseteq \mathbf{A}(x, n, F)$ and $\mathbf{B}_K(x, F_K) \supseteq \mathbf{B}(x, n, F)$ for all $n \geq N_K$ and for all $x \in \mathbf{X}$. In addition, for any history $\tilde{h}_{n-1}$

$$V(x, n, \tilde{h}_{n-1}u, \tilde{h}_{n-1}v) = (\beta_1)^{-n} \sum_{k=1}^{K} (\beta_k)^n V_k(x) \tag{28}$$

for $n \geq N_K$ and for any policies u and v that use actions from the sets $\mathbf{A}_K(x_t, F_K)$ and $\mathbf{B}_K(x_t, F_K)$ for all $t \geq N_K$. Therefore $\mathbf{A}(x, n, F) = \mathbf{A}_K(x, F_K)$ and $\mathbf{B}(x, n, F) = \mathbf{B}_K(x, F_K)$ for all $x \in \mathbf{X}$ and for all $n \geq N = N_K$. ∎

Corollary 5.1 *In a weighted discounted zero-sum stochastic game with finite state and action sets, for some $N < \infty$ each player has a persistently optimal (N, ∞)-stationary policy.*

Algorithm 5.1 0. Set $k = 1$.

1. Compute $V_k(x)$, $\mathbf{A}_k(x)$, and $\mathbf{B}_k(x)$ for all $x \in \mathbf{X}$. Compute N_k.

2. If $\mathbf{A}_k(x)$ and $\mathbf{B}_k(x)$ are singletons for all $x \in \mathbf{X}$ or $k = K$, set $\tilde{\mathbf{A}}(x) = \mathbf{A}_k(x)$ and $\tilde{\mathbf{B}}(x) = \mathbf{B}_k(x)$ all $x \in \mathbf{X}$ and continue to the next step. Otherwise increase k by one and repeat from step 1.

3. Fix stationary policies $\tilde{u}$ and $\tilde{v}$ for players 1 and 2 respectively, where $\tilde{u}(x) \in \tilde{\mathbf{A}}(x)$ and $\tilde{v}(x) \in \tilde{\mathbf{B}}(x)$ for all $x \in \mathbf{X}$.

4. Compute $F_N(x) = \sum_{k=1}^{K} (\beta_k)^N V_k(x, \tilde{u}, \tilde{v})$ for all $x \in \mathbf{X}$, where $N = N_K$.

5. Compute N-stage optimal pure Markov policies (u, v) by solving the N stage stochastic zero-sum game with perfect information with state space $\mathbf{X}$, action sets $\mathbf{A}(x)$ and $\mathbf{B}(x)$ for players 1 and 2 respectively, transition probabilities p, and rewards $r_t = \sum_{k=1}^{K} (\beta_k)^t r_k$. Since $\mathbf{A}(x)$ ($\mathbf{B}(x)$) are singletons for $x \in \mathbf{Z}$ ($x \in \mathbf{Y}$), $u_t(x)$ ($v_t(x)$) are defined in a unique way for $x \in \mathbf{Z}$ ($x \in \mathbf{Y}$), $t = 0, \ldots, N-1$. For $t = 0, \ldots, N-1$ the policies u, v can be defined by

$$F_t(x) = \max_{a \in \mathbf{A}(x)} \{ r_t(x, a, b) + \sum_{z \in \mathbf{X}} p(z|x, a, b) F_{t+1}(z) \} = r_t(x, u_t(x), b) + \sum_{z \in \mathbf{X}} p(z|x, u_t(x), b) F_{t+1}(z) \tag{29}$$

for $x \in \mathbf{Y}$ and by

$$F_t(x) = \min_{b \in \mathbf{B}(x)} \{ r_t(x, a, b) + \sum_{z \in \mathbf{X}} p(z|x, a, b) F_{t+1}(z) \} = r_t(x, a, v_t(x)) + \sum_{z \in \mathbf{X}} p(z|x, a, v_t(x)) F_{t+1}(z) \tag{30}$$

for $x \in \mathbf{Z}$.

6. (N, ∞)-stationary policies u and v for players 1 and 2 are optimal, where Step 5 defines $u_t(\cdot)$ and $v_t(\cdot)$ for $t = 0, \ldots, N-1$ and $u_t(\cdot) = \tilde{u}(\cdot)$, $v_t(\cdot) = \tilde{v}(\cdot)$ for $t \geq N$.

Remark 5.1 A minor modification of the algorithm leads to the computation of the sets of persistently optimal policies described in Theorems 4.2 and 5.1. The algorithm computes N, $\mathbf{A}(x, N, F_K) = \tilde{\mathbf{A}}(x)$, $\mathbf{B}(x, N, F_K) = \tilde{\mathbf{B}}(x)$. A minor modification of Step 5 leads to the computation of $\mathbf{A}(x, t, F_K)$ and $\mathbf{B}(x, t, F_K)$ as sets of actions at which maximums in (29) and minimums in (30) are attained, $t = 0, \ldots, N-1$.

Remark 5.2 The number N which the algorithm computes is an upper bound for the actual threshold after which optimal policies must take actions from the sets $A_K(x, F_K)$ and $B_K(x, F_K)$. In fact, it compares the loss over one step due to an action which is non-optimal for criterion $l = 1, \ldots, K-1$, to the maximum gain from the next step onward due to criteria $l+1, \ldots, K$. Since this gain decreases faster than the losses, after some step the one-step loss cannot be compensated by payoffs with smaller discount factors. The numbers $\delta_{n,l}$ provide an upper estimate for this compensation. It is possible to sharpen this estimate by using the difference between values of MDPs when both players maximize and minimize their payoffs. This approach was used in Algorithm 3.7 in Feinberg and Shwartz [4]. It provides a better upper estimate for N, but requires solutions of up to $K(K-1)$ additional MDPs.

6 Counterexamples

The first example describes a stochastic game with weighted discounted payoffs and with finite state and action sets in which there is no optimal policy which is stationary from some epoch onward. This shows that the perfect information structure is essential. The existence of ϵ-optimal policies with this property was proved in Filar and Vrieze [9].

Example 6.1 Consider a single state, which will be omitted from the notation below (we are thus in the framework of repeated games). Let $\mathbf{A} = \{1, 2\}, \mathbf{B} = \{1, 2\}$. Let $r_1(a, b) = 1\{a = b\}$, and $r_2(a, b) = 1\{b = 2\}$. Assume $1 > \beta_1 > \beta_2 > 0$, and define the total payoff:

$$V(u, v) = E^{(u,v)} \left[\sum_{t=0}^{\infty} \beta_1^t r_1(a_t, b_t) + \sum_{t=0}^{\infty} \beta_2^t r_2(a_t, b_t) \right].$$

Then the optimal policy for player 1 (that controls the actions $\mathbf{A}$) for all t large enough is to use action 1 with probability

$$(1 + [\beta_2/\beta_1]^t)/2. \tag{31}$$

This converges to a limit $1/2$, but not in finite time.

To obtain (31), we note that for any 2 by 2 matrix game R, for which a dominating policy does not exist for either player, the optimal policy is the one that results in the indifference to the other players policy. (This is true whether player 1 minimizes or maximizes). Hence, the optimal policy of player 1 in the matrix game, $u(1)$ and $u(2)$, satisfies

$$R_{11}u(1) + R_{21}u(2) = R_{12}u(1) + R_{22}u(2)$$

and hence

$$u(1) = \frac{R_{22} - R_{21}}{R_{11} - R_{12} - R_{21} + R_{22}}.$$

In our repeated game, R is the matrix

$$\beta_1^t \left\{ \left| \begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right| + \frac{\beta_2^t}{\beta_1^t} \left| \begin{array}{cc} 0 & 1 \\ 0 & 1 \end{array} \right| \right\}.$$

∎

In the next example, we consider a Markov Decision Process with one state, a compact set of actions, and continuous payoffs. Example 3.16 in Feinberg and Shwartz [4] shows that optimal (N, ∞)-stationary policies may not exist under these conditions. In the next example, there is a unique optimal Markov policy u, but the sequence u_t does not have a limit.

Example 6.2 Here again we assume a single state, and the actions are $\mathbf{A} = [-1, 1]$. Assume $1 > \beta_1 > \beta_2 > 0$. Let $r_1(a) = |a|$.

The single controller maximizes the expected total reward:

$$V(u) = E^u \left[\sum_{t=0}^{\infty} \beta_1^t r_1(a_t) + \sum_{t=0}^{\infty} \beta_2^t r_2(A_t) \right].$$

Let $r_2(a) = \sqrt{1 - |a|}$. If the component r_2 did not exist, the optimal actions would be $|a| = 1$. Now in the presence of r_2, since the derivative of r_2 at $a = -1$ (and $a = 1$) is ∞ ($-\infty$, resp.), the optimal policies u_t satisfy $|u_t| < 1$ for all t. Here again the convergence to the limit optimal policy $|a| = 1$ does not take finite time. Note that for any t, there are two optimizing actions.

Next we modify r_2 slightly so as to destroy the convergence. For positive a, we replace $r_2(a)$ with a linear interpolation of $\sqrt{1 - |a|}$ between points $a = 1 - (2n)^{-1}$, $n = 1, 2, \ldots$. For negative a, we replace $r_2(a)$ with a linear interpolation of $\sqrt{1 - |a|}$ between points $a = 1 - (2n+1)^{-1}$, $n = 1, 2, \ldots$. As a result, for each t there will be just one optimizing action, u_t. We have that $\lim_{t \to \infty} |u_t| = 1$, but u_t is going to be infinitely often close to -1, and infinitely often close to 1. Hence, it does not converge. (Note however that in the sense of set convergence $\limsup_{t \to \infty} \{u_t\} = \{-1, 1\}$). ∎

It is well known that non zero-sum games with perfect information need not have deterministic equilibria policies. This was illustrated by Federgruen in [7] section 6.6. A natural question is whether for weighted discounted stochastic games with perfect information, there exist equilibrium policies which are stationary from some epoch n onward.

Example 6.3 The following counterexample shows that the answer is negative. Consider a game with two players. Let the payoff function of player i be

$$W_i(x,u,v) = V_{i,1}(x,u,v) + V_{i,2}(x,u,v), \tag{32}$$

where

$$V_{ik} = E_x^{u,v} \sum_{t=0}^{\infty} (\beta_k)^t r_{ik}(x_t, a_t, b_t)$$

with $i, k = 1, 2$ and $\beta_1 > \beta_2$.

Let $X = \{1,2\}$, $A(1) = B(2) = \{1,2\}$, and $A(2) = B(1) = \{1\}$. We also have $p(1|1,1,1) = p(1|2,1,1) = \frac{2}{3}$ and $p(1|1,2,1) = p(1|2,1,2) = \frac{1}{3}$. The one-step rewards are $r_{21}(1,1,1) = r_{11}(2,1,1) = 1$, $r_{22}(1,2,1) = r_{22}(2,1,2) = -1$, $r_{12}(2,1,1) = r_{12}(2,1,2) = r_{22}(1,1,1) = r_{22}(1,2,1) = 1$ and all other rewards are 0. We remark that if we remove the second summand from (32), we get an example from Federgruen [7], section 6.6, in which there is no equilibrium deterministic policies for a standard discounted non-zero sum game with perfect information.

Any stationary policy u (v) of player 1 (2) is defined by a probability $p = u(1|1)$ $(q = v(1|2))$, $p, q \in [0,1]$. Let $P(p,q)$ be a matrix of transition probabilities of a Markov chain defined on X by a couple of policies (p,q),

$$P(p,q) = \begin{pmatrix} \dfrac{1+p}{3} & \dfrac{2+p}{3} \\ \\ \dfrac{1+q}{3} & \dfrac{2-q}{3} \end{pmatrix}.$$

A straightforward computation leads to

$$(I - \beta P(p,q))^{-1} = [(1-\beta)(3-\beta(p-q))]^{-1} \begin{pmatrix} 3-\beta(2-q) & \beta(2-p) \\ \\ \beta(1+q) & 3-\beta(1+p) \end{pmatrix}.$$

Let also $r_{ik}(p,q)$ be a one-step expected payoff vector if a couple of policies (p,q) is used,

$$r_{11}(p,q) = \begin{pmatrix} 0 \\ \\ 2q-1 \end{pmatrix}, \qquad r_{21}(p,q) = \begin{pmatrix} 2p-1 \\ \\ 0 \end{pmatrix},$$

$$r_{12}(p,q) = \begin{pmatrix} 0 \\ \\ 1 \end{pmatrix}, \qquad r_{12}(p,q) = \begin{pmatrix} 1 \\ \\ 0 \end{pmatrix}.$$

Let there exist an equilibrium policy which is stationary from some epoch N onward. Since all transition probabilities are positive, this means that there exists a couple (p^*, q^*) which is equilibrium for any objective vector $(W_1(x,n,p,q), W_2(x,n,p,q))$, $n = N, N+1, \ldots,$ $x \in X$,

$$W_i(x,n,p,q) = V_{i1}(x,p,q) + \left(\frac{\beta_2}{\beta_1}\right)^n V_{i2}(x,p,q),$$

where $i = 1, 2$.

We have that $W_i(x,n,p,q) \to V_{i1}(x,p,q)$ and therefore (p^*, q^*) is an equilibrium pair of policies for the standard discounted game with the payoff vector $(V_{11}(x,p,q), V_{21}(x,p,q))$. This game is described in section 6.6 of Federgruen [7] and it has a unique stationary equilibrium solution $p^* = q^* = .5$.

We also have that

$$W_1(1,n,p,q) = (2q-1) f(\beta_1, p, q) + \left(\frac{\beta_2}{\beta_1}\right)^n f(\beta_2, p, q),$$

$$W_2(1,n,p,q) = (2p-1) g(\beta_1, p, q) + \left(\frac{\beta_2}{\beta_1}\right)^n g(\beta_2, p, q),$$

where

$$f(\beta, p, q) = \frac{\beta(2-p)}{(1-\beta)(3-\beta(p-q))},$$

$$g(\beta, p, q) = \frac{(3-\beta(2-q))}{(1-\beta)(3-\beta(p-q))}.$$

We have that $\dfrac{\partial W_1(1,n,p,q)}{\partial p}|_{p=q=.5} < 0$ and $\dfrac{\partial W_2(1,n,p,q)}{\partial q}|_{p=q=.5} > 0$.

Therefore, there is no equilibrium policy which is stationary from some epoch n onward. ∎

Acknowledgment. Research of the second author was partially supported by NSF Grant DMI-9500746. Research of the third author was supported in part by the Israel Science Foundation, administered by the Israel Academy of Sciences and Humanities, and in part by the fund for promotion of research at the Technion.

References

[1] E. Altman, A. Hordijk and F. M. Spieksma, "Contraction Conditions for Average and α-Discounted Optimality in Countable State Markov Games with Unbounded Payoffs", *MOR*, **22** No. 3, pp. 588-618, 1997.

[2] P. Billingsley, *Convergence of Probability Measures.* John Wiley, New York, 1968.

[3] E. A. Feinberg "Controlled Markov Processes with Arbitrary Numerical Criteria" *Theory Probab. Appl.* **27**, pp. 486-503, 1982.

[4] E. A. Feinberg and A. Shwartz, "Markov Decision Models with Weighted Discounted Criteria", *Math. Oper. Res.* **19**, pp. 152-168, 1994.

[5] E. A. Feinberg and A. Shwartz, "Constrained Markov Decision Models with Weighted Discounted Criteria", *Math. Oper. Res.* **20**, pp. 302-320, 1994.

[6] E. A. Feinberg and A. Shwartz, "Constrained Dynamic Programming with Two Discount Factors: Applications and an Algorithm," to appear in *IEEE Trans. Autom. Contr.*

[7] A. Federgruen, "On N-person Stochastic Games with Denumerable State Space", *Adv. Appl. Prob.* **10**, pp. 452-471, 1978.

[8] E. Fernández-Gaucherand, M. K. Ghosh and S. I. Marcus "Controlled Markov Processes on the Infinite Planning Horizon: Weighted and Overtaking Cost Criteria," *ZOR – Methods and Models of Operations Research* **39**, pp. 131-155, 1994.

[9] J. A. Filar and O. J. Vrieze, "Weighted Reward Criteria in Competitive Markov Decision Processes", *ZOR – Methods and Models of Operations Research* **36**, pp. 343-358, 1992.

[10] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*, Springer, New York, 1996.

[11] D. Gillette, "Stochastic Games with Zero Stop Probabilities", *Contribution to the Theory of Games*, III, M. Dresher, A. W. Tucker, P. Wolfe, eds., Princeton University Press, Princeton, 1957, pp. 179-187.

[12] K. Golabi, R. B. Kulkarni and G. B. Way, "A statewide Pavement Management System," *Interfaces*, **12**, pp. 5-21, 1982.

[13] C. J. Himmelberg, T. Parthasarathy, T. E. S. Raghavan, "Existence of p-equilibrium and Optimal Stationary Strategies in Stochastic Games", *Proc. Amer. Math. Soc.* **60**, pp. 245-251, 1976.

[14] S. Karlin, *Mathematical Methods and Theory in Games, Programming, and Economics. Volume II: The Theory of Infinite Games*, Addison-Wesley, New York, 1959

[15] D. Krass, J. A. Filar and S. S. Sinha, "A Weighted Markov Decision Process", *Oper. Res.* **40**, pp. 1180-1187, 1992.

[16] H.-U. Küenle, *Stochastiche Spiele und Entscheidungesmodelle*, Tebuner-Texte, Band 89, 1986.

[17] P.R. Kumar and T. H. Shiau, "Existence of Value and Randomized Strategies in Zero-Sum Discrete-Time Stochastic Dynamic Games", *SIAM J. Cont. Opt.* **19**, pp. 617-634, 1981.

[18] A. Maitra and W. D. Sudderth, *Discrete Gambling and Stochastic Games*, Springer, New York, 1996.

[19] A. S. Nowak, "On Zero-Sum Stochastic Games with General State Space I", *Probab. Math. Statist.* **4**, pp. 13-32, 1984.

[20] A. S. Nowak, "Universally Measurable Strategies in Stochastic Games", *Ann. Probab.* **13**, pp. 269-287, 1985.

[21] A. S. Nowak, "Semicontinuous Nonstationary Stochastic Games", *J. Math. Anal. Appl.* **117**, pp. 84-99, 1986.

[22] K. R. Parthasarathy, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.

[23] T. Parthasarathy and E. S. Raghavan, *Some Topics in Two-Person Games*, Elsevier, New York, 1971.

[24] M. I. Reiman and A. Shwartz, "Call Admission: a New Approach to Quality of Service," CC Pub. 216, Technion, and Bell Labs Manuscript, 1997.

[25] L. S. Shapley (1953), "Stochastic Games", *Proc. Nat. Acad. Sci. USA* **39**, pp. 1095-1100.

[26] L. S. Shapley and R. N. Snow, "Basic Solutions of Discrete Games", *Contribution to the Theory of Games*, I, H. W. Kuhn, A. W. Tucker, eds., Princeton University Press, Princeton, 1957, pp. 27-35.

[27] M. Sion, "On General Minimax Theorems", *Pacific J. Math.* **8**, pp. 171-176, 1958.

[28] C. V. Winden and R. Dekker, "Markov decision models for building maintenance: a feasibility study," Report 9473/A, ERASMUS University Rotterdam, The Netherlands, 1994.