Stationary Anonymous Sequential Games with Undiscounted Rewards

Piotr Wiecek, Eitan Altman

Stationary Anonymous Sequential Games with Undiscounted Rewards

Piotr Więcek · Eitan Altman

Abstract Stationary anonymous sequential games with undiscounted rewards are a special class of games that combines features from both population games (infinitely many players) with stochastic games. We extend the theory for these games to the cases of total expected reward as well as to the expected average reward. We show that equilibria in the anonymous sequential game correspond to the limits of equilibria of related finite population games as the number of players grows to infinity. We provide examples to illustrate our results.

Keywords Stochastic game · Population game · Anonymous sequential game · Average reward · Total reward · Stationary policy

1 Introduction

Games with a continuum of atomless (or infinitesimal) players have since long ago been used to model interactions involving a large number of players in which the action of a single player has a negligible impact on the utilities of other players. In road traffic engineering, for example, this was already formalized by Wardrop [1] in 1952 to model the choice of routes of cars where each driver, modeled as an atomless player, minimizes its expected travel delay. In Wardrop's model, there may be several classes of players, each corresponding to another origin-destination pair. The goal is to determine what fraction of each class of players would use the different possible paths available to that

This work is supported by the NCN Grant no DEC- 2011/03/B/ST1/00325.

P. Więcek
Institute of Mathematics and Computer Science, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
Tel.: +48-071-320-31-60
Fax: +48-071-328-07-51
E-mail: Piotr.Wiecek@pwr.wroc.pl

E. Altman
INRIA, 2004 Route des Lucioles, P.B. 93, 06902 Sophia Antipolis Cedex, France

class. The equilibrium is known to behave as the limit of the equilibria obtained in games with finitely many players, as their number tends to infinity [2]. It is also the limit of Nash equilibria for some sequence of dynamic games in which randomness tends to average away as the number of players increases [3].

Another class of games that involves a continuum of atomless players are evolutionary games, in which pairs of players that play a matrix game are selected at random, see [4]. Our objective is again to predict the fraction of the population (or of populations in the case of several classes) that play each possible action at equilibrium. A Wardrop type definition of equilibrium can be used, although there has been a particular interest in a more robust notion of equilibrium strategy, called Evolutionary Stable Strategy (we refer the reader to [5,6]).

In both games described above, the player's type is fixed, and the actions of the players determine directly their utilities.

Extensions of these models are needed whenever the player's class may change randomly in time, and when the utility of a player depends not only on the current actions of players but also on future interactions. The class of the player is called its individual state. The choice of an action by a player should then take into account not only the game played at the present state but the future state evolution. We are interested in particular in the case where the action of a player not only impacts the current utility but also the transition probabilities to the next state.

In this paper we study this type of extension in the framework of the first type of game, in which a player interacts with an infinite number of other players. (In the road traffic context, the interaction is modeled through link delays each of which depends on the total amount of traffic that uses that link.) We build upon the framework of anonymous sequential games, introduced by B. Jovanovic and R.W. Rosenthal in 1988 in [7]. In that work, each player's utility is given as the expected discounted utility over an infinite horizon. The theory of anonymous sequential games with discounted utilities was further developped in [8–12]. Conditions under which Nash equilibria in finite-player discounted-utility games converge to equilibria of respective anonymous models were analyzed in [13–16]. Also applications of this kind of models were numerous: from stochastic growth [17] and industry dynamics [18–21] models to dynamic auctions [22–24] and strategic market games [25,26]. Surprisingly, the cases of expected average utility and total expected utility have remained open ever since 1988, even though this kind of models were applied in some networking contexts [27,28]. Our main contribution in this paper is giving conditions under which such extensions are possible.

Similar extensions have been proposed and studied for the framework of evolutionary games in [29,30]. The analysis there turns out to be simpler since the utility in each encounter between two players turns out to be bilinear there.

The structure of the paper is as follows. We begin with a section that presents the model and introduces in particular the expected average and the total expected reward criteria. The two following sections establish the

existence of stationary equilibria for the average and the total reward (Section 3 and 4, respectively). Section 5 is concerned with showing that the equilibria for models of the two previous chapters that deal with infinite number of players are limits of those obtained for some games with a large finite number of players, as this number goes to infinity. We end with two sections that show how our results apply to some real-life examples, followed by a concluding paragraph.

2 The Model

The anonymous sequential game is described by the following objects:

- We assume that the game is played in discrete time, that is $t \in \{1, 2, \ldots\}$.
- The game is played by an infinite number (continuum) of players. Each player has his own private state $s \in S$, changing over time. We assume that S is a finite set.
- The global state, μ^t, of the system at time t, is a probability distribution over S. It describes the proportion of the population, which is at time t in each of the individual states. We assume that each player has an ability to observe the global state of the game, so from his point of view the state of the game at time t is[1] $(s_t, \mu^t) \in S \times \Delta(S)$.
- The set of actions available to a player in state (s, μ) is a nonempty set $A(s, \mu)$, with $A := \bigcup_{(s,\mu)\in S\times\Delta(S)} A(s, \mu)$ – a finite set. We assume that the mapping A is an upper semicontinuous function.
- Global distribution of the state-action pairs at any time t is given by the measure $\tau^t \in \Delta(S \times A)$. The global state of the system μ^t is the marginal of τ^t on S.
- An individual's immediate reward at any stage t, when his private state is s_t, he plays action a_t and the global state-action measure is τ^t is $u(s_t, a_t, \tau^t)$. It is a (jointly) continuous function.
- The transitions are defined for each individual separately with the transition function $Q : S \times A \times \Delta(S \times A) \to \Delta(S)$ which is also a (jointly) continuous function. We will write $Q(\cdot|s_t, a_t, \tau^t)$ for the distribution of the individual state at time $t + 1$, given his state at time t, s_t, his action a_t and the state-action distribution of all the players.
- The global state at time $t+1$ will be given by[2] $\Phi(\cdot|\tau^t) = \sum_{s\in S}\sum_{a\in A} Q(\cdot|s, a, \tau^t)\tau^t_{sa}$.

Any function $f : S \times \Delta(S) \to \Delta(A)$ satisfying $\mathrm{supp} f(s, \mu) \subset A(s, \mu)$ for every $s \in S$ and $\mu \in \Delta(S)$ is called a *stationary policy*. We denote the set of stationary policies in our game by $\mathcal{U}$.

[1] Here and in the sequel for any set B, $\Delta(B)$ denotes the set of all the finite-support probability measures on B. In particular, if B is a finite set, it denotes the set of all the probability measures over B. In such a case we always assume that $\Delta(B)$ is endowed with Euclidean topology.

[2] Note that its transition is deterministic.

2.1 Average reward

We define the *long-time average reward* of a player using stationary policy f when all the other players use policy g and the initial state distribution (both of the player and his opponents) is μ^1, to be

$$J(\mu^1, f, g) = \limsup_{T\to\infty} \frac{1}{T} E^{\mu^1,Q,f,g} \sum_{t=1}^{T} u(s_t, a_t, \tau^t).$$

Further, we define a stationary strategy f and a measure $\mu \in \Delta(S)$ to be an equilibrium in the long-time average reward game if for every other stationary strategy $g \in \mathcal{U}$,

$$J(\mu, f, f) \geq J(\mu, g, f)$$

and, if $\mu^1 = \mu$ and all the players use policy f then $\mu^t = \mu$ for every $t \geq 1$.

Remark 1 The definition of the equilibrium used here differs significantly from that used in [7]. There the equilibrium is defined with respect to the solution of some dynamic programming. Our definition directly relates it to the reward functionals.

2.2 Total reward

To define the total reward in our game let us distinguish one state in S, say s_0 and assume that $A(s_0, \mu) = \{a_0\}$ independently of μ for some fixed a_0. Then the *total reward* of a player using stationary policy f when all the other players apply policy g and the initial distribution of the states of his opponents is μ^1, while his own is ρ^1, is defined in the following way:

$$\overline{J}(\rho^1, \mu^1, f, g) = E^{\rho^1,\mu^1,Q,f,g} \sum_{t=1}^{\mathcal{T}-1} u(s_t, a_t, \tau^t),$$

where $\mathcal{T}$ is the moment of the first arrival of the process s_t to s_0. We interpret it as the reward accumulated by the player over whole of his lifetime. State s_0 is an artificial state (so is action a_0) denoting that a player is dead. μ^1 is the distribution of the states across the population when he is born, while ρ^1 is the distribution of initial states of new-born players. The fact that after some time the state of a player can become again different from s_0 should be interpreted as that after some time the player is replaced by some new-born one.

The notion of equilibrium for the total reward case will be slightly different from that for the average reward. We define a stationary strategy f and a measure $\mu \in \Delta(S)$ to be in equilibrium in the total reward game if for every other stationary strategy $g \in \mathcal{U}$,

$$\overline{J}(\rho, \mu, f, f) \geq \overline{J}(\rho, \mu, g, f),$$

where $\rho = Q(\cdot|s_0, a_0, \tau(f, \mu))$ and $(\tau(f, \mu))_{sa} = \mu_s(f(s))_a$ for all $s \in S$, $a \in A$, and, if $\mu^1 = \mu$ and all the players use policy f then $\mu^t = \mu$ for every $t \geq 1$.

3 Existence of the Stationary Equilibrium in Average-reward Case

In the present section we present a result about the existence of stationary equilibrium in anonymous sequential games with long-time average reward. We prove it under the following assumption:

(A1) The set of individual states of any player S can be partitioned into two sets S_0 and S_1 such that for every state-action distribution of all the other players $\tau \in \Delta(S \times A)$:
- (a) All the states from S_0 are transient in the Markov chain of individual states of a player using any $f \in \mathcal{U}$.
- (b) The set S_1 is strongly communicating.

There are a couple of equivalent definitions of "strongly communicating" property used above appearing in the literature. We follow the one formulated in [31], saying that a set S_1 of states in a Markov Decision Process is strongly communicating if there exists a stationary policy[3] $\overline{f}^{\tau}$ such that the probability of going from any state $s \in S_1$ to any other $s' \in S_1$ for a player using $\overline{f}^{\tau}$ is positive.

Assumption (A1) appears often in the literature on Markov decision processes with average cost and is referred to as "weakly communicating" property, see e.g. [32], chapters 8 and 9.[4] It guarantees that the optimal gain in a Markov decision process satisfying it is independent of its initial state. As we will see, it also guarantees that this optimal gain is continuous in τ, which will be crucial in proving the existence of an equilibrium in our game. It is also worth noting that without assumption (A1) the average-reward anonymous sequential game may have no stationary equilibria at all. This is shown in the following example[5]

Example 1 Let us consider an average reward anonymous sequential game with $S = \{1, 2, 3\}$ and $A(s, \mu) = \begin{cases} \{0, 1\}, & \text{if } s = 1 \\ \{0\}, & \text{otherwise} \end{cases}$, thus the decision is only made by players in state 1. For the simplicity we will denote this only decision by a in what will follow. The immediate rewards for the players depend only on their private state as follows: $u(s) = 3 - s$. Finally, the transition matrix of the Markov chain of private states of each player is

$$\mathbb{Q}(a, \tau) = \begin{bmatrix} 1 - \frac{a+3p^*}{4} & \frac{a}{4} & \frac{3p^*}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{p^*}{2} & 0 & 1 - \frac{p^*}{2} \end{bmatrix}, \text{ where } p^* = \max\{0, 1 - 4\tau_{11}\}.$$

[3] Note that we assume this stationary policy may depend on τ, as we consider the properties of the Markov chain of individual states of a player under fixed state-action distribution of all the other players.

[4] All the properties appearing in the assumptions are commonly used in the Markov decision processes literature. Those readers who are not familiar with them or are interested in mutual relationships between these properties are referred to [31,33].

[5] The example is a reworking of Example 3. in [34].

It violates assumption (A1), as e.g. when $\tau_{11} \geq \frac{1}{4}$ for the pure strategy assigning $a = 0$ in state 1, states 1 and 2 are absorbing in the Markov chain of individual states of a player and state 2 is transient, while when it assigns $a = 1$ states 1 and 2 become communicating. We will show that such a game has no stationary equilibrium.

Suppose that (f, μ^*) is an equilibrium. We will consider two cases:

(a) $\tau_{11} \geq \frac{1}{4}$ for τ corresponding to μ^* and f. Then $p^* = 0$, and so if a player uses action 1 with probability β, the stationary state of the chain of his states when his initial state's distribution is μ^* is $\left[\frac{2(\mu_1^*+\mu_2^*)}{2+\beta}, \frac{\beta(\mu_1^*+\mu_2^*)}{2+\beta}, \mu_3^*\right]$ and his long-time average reward is

$$\frac{(4+\beta)(\mu_1^*+\mu_2^*)}{2+\beta} = \left(1 + \frac{2}{2+\beta}\right)(\mu_1^*+\mu_2^*),$$

which is a strictly decreasing function of β (recall that $\mu_1^* \geq \tau_{11} \geq \frac{1}{4}$). Thus his best response to f is the policy which assigns probability 1 to action $a = 0$ in state 1. But if all the players use such policy, $\tau_{11} = 0$, which contradicts our assumption that it is no less than $\frac{1}{4}$.

(b) $\tau_{11} < \frac{1}{4}$. Then it can be easily seen that the stationary state of any player's chain when he uses action 1 with probability $\beta \in [0, 1]$ is independent of the initial distribution of his state μ^* and equal to $\left[\frac{2}{5+\beta}, \frac{\beta}{5+\beta}, \frac{3}{5+\beta}\right]$, which gives him the reward of $\frac{4+\beta}{5+\beta} = 1 - \frac{1}{5+\beta}$, which is clearly a strictly increasing function of β. Thus the best response to f is to play action $a = 1$ with probability 1, which, if applied by all the players, results in stationary state $\mu^* = \left[\frac{1}{3}, \frac{1}{6}, \frac{1}{2}\right]$ and consequently $\tau_{11} = \frac{1}{3}$, contradicting the assumption that it is less than $\frac{1}{4}$.

Thus this game cannot have a stationary equilibrium.

Now we are ready to formulate the main result of this section.

Theorem 1 *Every anonymous sequential game with long-time average payoff satisfying (A1) has a stationary equilibrium.*

Before we prove the theorem let us introduce some additional notation. We will consider a Markov decision process $\mathcal{M}(\tau)$ of an individual faced with a fixed (over time) distribution of state-action pairs of all the other players. For this fixed $\tau \in \Delta(S \times A)$ let $J^\tau(f, \mu)$ denote the long-time average payoff in this process when the player uses stationary policy f and the initial distribution of states is μ, that is

$$J^\tau(f, \mu) = \limsup_{T \to \infty} \frac{1}{T} E^{\mu, Q, f} \sum_{t=1}^{T} u(s_t, a_t, \tau).$$

By well known results from dynamic programming (see e.g. [32]), in a weakly communicating Markov decision process (this is such a process by (A1)) the

optimal gain is independent of μ. We denote this uniform optimal gain by $G(\tau)$, that is

$$G(\tau) = \sup_{f:S\to A} J^\tau(f,\mu) \quad \text{for any fixed } \mu \in \Delta(S).$$

Lemma 2 states a crucial feature of G. It is preceded by another technical one.

Lemma 1 *Suppose that $\mu(n)$ are invariant measures of Markov chains with finite state set S and with transition matrices $P(n)$ respectively. Then if $\mu(n) \to \mu$ and $P(n) \to P$, then μ is an invariant measure for the Markov chain with transition matrix P.*

Proof: By the definition of invariant measure every $\mu(n)$ satisfies for every $s \in S$

$$(\mu(n))_s = \sum_{i\in S} (\mu(n))_i (P(n))_{si}.$$

If we pass to the limit, we obtain

$$\mu_s = \sum_{i\in S} \mu_i P_{si}.,$$

which means that μ is an invariant measure for the Markov chain with transition matrix P. □

Lemma 2 *Under (A1) G is a continuous function of τ.*

Proof: Let τ^n be a sequence of probability measures on $S \times A$ converging to τ. Since all of the MDPs we consider here have finite state space, each of them has a stationary optimal policy[6], say policy f^n is optimal in $\mathcal{M}(\tau^n)$. Next, let μ^n be an invariant measure corresponding to strategy f^n in $\mathcal{M}(\tau^n)$ (by (A1) such a measure exists, maybe more than one). Such an invariant measure must satisfy

$$G(\tau^n) = \sum_{s\in S}\sum_{a\in A} (f^n(s))_a \mu^n_s u(s,a,\tau^n), \tag{1}$$

otherwise f^n would not be optimal.

Note next that $\mathcal{U}$ and $\Delta(S)$ are compact sets. Thus there exists a subsequence of f^n converging to some $f^0 \in \mathcal{U}$ and then a subsequence of μ^n converging to some μ^0. Without a loss of generality we may assume that these are sequences f^n and μ^n that are convergent. By the continuity assumption about Q and Lemma 1, μ^0 is an invariant measure corresponding to f^0 in $\mathcal{M}(\tau)$. If we next pass to the limit in (1) we get

$$\lim_{n\to\infty} G(\tau^n) = \sum_{s\in S}\sum_{a\in A} (f^0(s))_a \mu^0_s u(s,a,\tau) = J^\tau(f^0,\mu^0).$$

But this implies that $G(\tau) \geq \lim_{n\to\infty} G(\tau^n)$.

[6] Of course a stationary policy is only a function of individual state s here.

To show the inverse inequality, suppose that $f \in \mathcal{U}$ is an optimal policy in $\mathcal{M}(\tau)$. By (A1) the states in the Markov chain of individual states for a user applying f in $\mathcal{M}(\tau)$ can be divided into a class of transient states and a number of communicating classes. Let S^* be a communicating class such that the ergodic payoff in this class is equal to $G(\tau)$. Define now the policies g^n in the following way:

$$g^n(s) = \begin{cases} f(s) & \text{when } s \in S^* \\ f^{\tau^n}(s) & \text{when } s \in S \setminus S^* \end{cases}$$

(Here f^{τ^n} is a communicating policy derived from assumption (A1)). One can easily notice that under these policies applied in $\mathcal{M}(\tau^n)$, all the states from $S \setminus S^*$ would be transient. Now let $\overline{\mu}^n$ be an invariant measure corresponding to g^n in $\mathcal{M}(\tau^n)$. Again using Lemma 1 we can show that the limit (possibly over a subsequence) of the sequence $\overline{\mu}^n$, say $\overline{\mu}^0$ is an invariant measure of the limit of g^n, which is equal to f on S^*. At the same time $\overline{\mu}^0_s = 0$ for $s \in S \setminus S^*$, so we can write (from the definition of invariant measure) for every $s \in S^*$:

$$\begin{aligned} \overline{\mu}^0_s &= \sum_{i \in S} \sum_{a \in A} Q(s|i,a,\tau)(g^0(i))_a \overline{\mu}^0_i = \sum_{i \in S^*} \sum_{a \in A} Q(s|i,a,\tau)(g^0(i))_a \overline{\mu}^0_i \\ &= \sum_{i \in S^*} \sum_{a \in A} Q(s|i,a,\tau)(f(i))_a \overline{\mu}^0_i = \sum_{i \in S} \sum_{a \in A} Q(s|i,a,\tau)(f(i))_a \overline{\mu}^0_i, \end{aligned}$$

which means that $\overline{\mu}^0$ is also an invariant measure for f and it is entirely concentrated on S^*. But since S^* is a communicating class under f, this implies

$$J^\tau(g^0, \overline{\mu}^0) = J^\tau(f, \overline{\mu}^0) = G(\tau).$$

On the other hand

$$\begin{aligned} J^\tau(g^0, \overline{\mu}^0) &= \sum_{s \in S} \sum_{a \in A} (g^0(s))_a \overline{\mu}^0_s u(s,a,\tau) \\ &= \lim_{n \to \infty} \sum_{s \in S} \sum_{a \in A} (g^n(s))_a \overline{\mu}^n_s u(s,a,\tau^n) = \lim_{n \to \infty} J^{\tau^n}(g^n, \overline{\mu}^n) \leq \lim_{n \to \infty} G(\tau^n), \end{aligned}$$

ending the proof. □

Proof of Theorem 1: We consider two multifunctions of $\tau \in \Delta(S \times A)$:

$$B(\tau) := \left\{ \rho \in \Delta(S \times A) : \sum_{s \in S} \sum_{a \in A} \rho_{sa} u(s,a,\tau) = G(\tau) \right\} \tag{2}$$

$$C(\tau) = \left\{ \rho \in \Delta(S \times A) : \sum_{a \in A} \rho_{sa} = \sum_{x \in S} \sum_{b \in B} Q(s|x,b,\tau) \rho_{xb} \right\} \tag{3}$$

and let $\Psi(\tau) := B(\tau) \cap C(\tau)$. We will show that Ψ has a fixed point, and then that this fixed point corresponds to an equilibrium in the game.

First note that $C(\tau)$ is the set of all the possible stationary state-action measures in $\mathcal{M}(\tau)$. By Theorem 1 in [35] it is also the set of occupation measures corresponding to all the possible stationary policies and all the possible initial distributions of states in $\mathcal{M}(\tau)$. Since $G(\tau)$ is the optimal reward in this MDP, there exists a stationary policy, and thus an occupation measure corresponding to it, for which the reward is equal to $G(\tau)$. This implies that for any τ, $\Psi(\tau)$ is nonempty. Further note that for any τ both $B(\tau)$ and $C(\tau)$ are trivially convex, and thus so is their intersection. Finally, as an immediate consequence of Lemma 1, C has a closed graph. On the other hand the closedness of the graph of B is a trivial consequence of the continuity of u (by assumption) and G (by Lemma 2). The graph of Ψ is the intersection of the two and thus is also closed. The existence of a fixed point of Ψ follows now from Glickberg's fixed point theorem [36].

Now suppose τ^* is this fixed point. Since it is a fixed point of C, it satisfies:

$$\sum_{a\in A}\tau^*_{sa} = \sum_{x\in S}\sum_{b\in B} Q(s|x,b,\tau^*)\tau^*_{xb} = \Phi(s|\tau^*).$$

This implies that if the initial distribution of states is τ^*_S and players apply stationary policy f defined for any $\tau \in \Delta(S\times A)$ by[7]:

$$(f(s),\tau)_a = \begin{cases} \frac{\tau^*_{sa}}{\sum_{b\in A}\tau^*_{sa}} & \text{if } \sum_{b\in A}\tau^*_{sa} > 0 \\ \delta[a_0] & \text{otherwise} \end{cases} \tag{4}$$

for any fixed $a_0 \in A$, the distribution of state-action pairs in the population is always τ^*. On the other hand, since $\tau^* \in B(\tau^*)$, f is the best response of a player when the state-action distribution is always τ^*, and thus together with τ^*_S an equilibrium in the game. □

4 Existence of the Stationary Equilibrium in Total-reward Case

In this section we show that also for the total reward case under some fairly mild assumptions the game has an equilibrium. What we will assume is the following:

(T1) There exists a $p_0 > 0$ such that for any fixed state-action measure τ and under any stationary policy f the probability of getting from any state $s \in S\setminus\{s_0\}$ to s_0 in $|S|-1$ steps is not smaller than p_0.

We write that this assumption is fairly mild, as it is not only necessary for our theorem to hold, but also for the total cost model to make sense, as it is trivially shown in an example below:

Example 2 Consider an anonymous sequential game with total cost with $S = \{1,2,3\}$ (and state s_0 denoted here as $s=0$) and $A(s,\mu) = \begin{cases} \{1,2\}, & \text{if } s=1 \\ \{1\}, & \text{otherwise} \end{cases}$.

[7] Here and in the sequel $\delta[x]$ denotes a probability measure concentrated in x.

The immediate rewards for the players depend only on their private state as follows: $u(s) = \begin{cases} 1, & \text{if } s = 1 \\ -1, & \text{otherwise} \end{cases}$. Finally, the transitions of the Markov chain of private states of each player are defined as follows: if his action in state 1 is 1, he moves with probability 1 to state 2; if his action is 2, he moves with probability 1 to state 3. In state 2 he moves to 1 with probability 1, while in state 3 he dies with probability $\frac{1}{2}$ and stays in 3 with probability $\frac{1}{2}$. This game is completely decoupled (in the sense that neither the rewards nor the transitions of any player depend on those of the others), so it is easy to analyze.

It is immediate to see that under pure stationary policy choosing action 1 in state 1 a player never dies. Moreover, he receives payoffs of 1 and -1 in subsequent periods, so his reward (the sum of these rewards over his lifetime) is not well defined. If we try to correct it by defining the total reward as the lim sup or lim inf of his accumulated rewards after n periods of his life, we obtain a strange situation that the policy choosing action 1 is optimal for each of the players in the lim sup version of the reward, and his worst possible policy for the lim inf version.

Remark 2 The total reward model, specifically when (T1) is assumed, bears a lot of resemblance to an exponentially discounted model where the discount factor is allowed to fluctuate over time, which suggests that the results in the two models should not differ much. Note however that there is one essential difference between these two models. The 'discount factor' in the total reward model (which is the ratio of those who stay alive after a given period to those who were alive at its beginning) appears not only in the cumulative reward of the players but also in the stationary state of the game, and thus also in the per-period rewards of the players. Thus this is an essentially different (and slightly more complex) problem. On the other hand, the fact that each of the players lives for a finite period and then is replaced by another player, with a fixed fraction of players dead and fixed fractions of players in each of the states when the game is in a stationary state, makes this model similiar to the average reward one. In fact, using the renewal theorem, we can relate the rewards of the players in the total reward model with those in the respective average reward model. This relation in used a couple of times in our proofs.

Now we can formulate our main result of this section.

Theorem 2 *Every anonymous sequential game with total reward satisfying (T1) has a stationary equilibrium.*

As in the case of the average reward, we start by defining some additional notation. Let $\overline{\mathcal{M}}(\tau)$ be a modified Markov decision process of an individual faced with a fixed (over time) distribution of state-action pairs of all the other players. This modification is slight but important, namely we assume that in $\overline{\mathcal{M}}(\tau)$ the state s_0 is absorbing and the reward in this state is always 0. This is a classic MDP with total reward, as considered in the literature. For this fixed $\tau \in \Delta(S \times A)$ let $\overline{J}^{\tau}(f, \rho)$ denote the total payoff in this process when

the player uses stationary policy f and the initial distribution of states is ρ, and let $\overline{G}(\tau)$ denote the optimal reward in $\overline{\mathcal{M}}(\tau)$, that is

$$\overline{G}(\tau) = \sup_{f:S\to A} \overline{J}^{\tau}(f, Q(\cdot|s_0, a_0, \tau)).$$

We can prove the following auxiliary result:

Lemma 3 *Under (T1) $\overline{J}^{\tau}(f,\rho)$ is a (jointly) continuous function of ρ, f and τ.*

Proof: By a well known result from the theory of Markov decision processes (see e.g. [32], Lemma 7.1.8 – assumption of this lemma is satisfied as a consequence of assumption (T1) and the fact that function u is continuous on a compact set and thus bounded), $\overline{J}^{\tau}(f,\rho)$ is the limit of the rewards in respective discounted MDPs $\overline{J}^{\tau}_{\beta}(f,\rho)$ as the discount factor β approaches 1. Since by a well known result (see e.g. Lemma 8.5 in [37]) discounted rewards are continuous functions of stationary policies and they are linear in the initial distribution of states, to show the continuity of $\overline{J}^{\tau}(f,\rho)$ it is enough to prove that the convergence of $\overline{J}^{\tau}_{\beta}(f,\rho)$ is uniform.

Let us take an $\varepsilon > 0$ and set $M = \max_{s,a,\tau} |u(s,a,\tau)|$ (such a number exists as u is a continuous function defined on a compact set). The probability that after $m(|S|-1)$ steps the state of an individual did not reach s_0 is for any stationary policy f by assumption (T1) not greater than $(1-p_0)^m$. This means that for any m,

$$\left| \overline{J}^{\tau}(f,\rho) - \sum_{t=1}^{m(|S|-1)} E^{\rho,Q,f} u(s_t,a_t,\tau) \right| \leq M(|S|-1)(1-p_0)^m, \tag{5}$$

which for m big enough, say bigger than m_ε, is not greater than $\frac{\varepsilon}{3}$. Note that analogously for $m \geq m_\varepsilon$ and any $\beta \in (0,1)$

$$\left| \overline{J}^{\tau}_{\beta}(f,\rho) - \sum_{t=1}^{m(|S|-1)} E^{\rho,Q,f} \beta^{t-1} u(s_t,a_t,\tau) \right| < \frac{\varepsilon}{3}, \tag{6}$$

Next note that

$$\left| \sum_{t=1}^{m(|S|-1)} E^{\rho,Q,f} u(s_t,a_t,\tau) - \sum_{t=1}^{m(|S|-1)} E^{\rho,Q,f} \beta^{t-1} u(s_t,a_t,\tau) \right| \leq M(1-\beta^{m(|S|-1)}) \leq \frac{\varepsilon}{3} \tag{7}$$

for β big enough, say $\beta > \beta_\varepsilon$. Combining (5), (6) and (7) and using the triangle inequality we obtain:

$$\left| \overline{J}^{\tau}(f,\rho) - \overline{J}^{\tau}_{\beta}(f,\rho) \right| \leq \left| \overline{J}^{\tau}(f,\rho) - \sum_{t=1}^{m(|S|-1)} E^{\rho,Q,f} u(s_t,a_t,\tau) \right|$$

$$+\left|\sum_{t=1}^{m(|S|-1)} E^{\rho,Q,f}u(s_t,a_t,\tau)-\sum_{t=1}^{m(|S|-1)} E^{\rho,Q,f}\beta^{t-1}u(s_t,a_t,\tau)\right|$$

$$+\left|\sum_{t=1}^{m(|S|-1)} E^{\rho,Q,f}\beta^{t-1}u(s_t,a_t,\tau)-\overline{J}^{\tau}_{\beta}(f,\rho)\right|\leq\frac{\varepsilon}{3}+\frac{\varepsilon}{3}+\frac{\varepsilon}{3}=\varepsilon$$

for $\beta>\beta_{\varepsilon}$, which ends the proof. □

Proof of Theorem 2: We shall consider two multifunctions of $\tau\in\Delta(S\times A)$:

$$\overline{B}(\tau):=\Bigg\{\rho\in\Delta(S\times A):\exists f^{\rho}\in\mathcal{U}$$
$$\forall s\in S,\left(\sum_{a\in A}\rho_{sa}>0\Rightarrow f^{\rho}(s)=\frac{\rho_{sa}}{\sum_{a\in A}\rho_{sa}}\right)$$
$$\text{and }\overline{J}^{\tau}(f^{\rho},Q(\cdot|s_0,a_0,\tau))=\overline{G}(\tau)\Bigg\}$$

$$\overline{C}(\tau)=\left\{\rho\in\Delta(S\times A):\sum_{a\in A}\rho_{sa}=\sum_{x\in S}\sum_{b\in B}Q(s|x,b,\tau)\rho_{xb}\right\}$$

and let $\overline{\Psi}(\tau):=\overline{B}(\tau)\cap\overline{C}(\tau)$. We will show that $\overline{\Psi}$ has a fixed point and that it corresponds to an equilibrium in the game.

First note that $\overline{\Psi}(\tau)$ is nonempty for any $\tau\in\Delta(S\times A)$, as any invariant measure corresponding to an optimal stationary policy in $\overline{\mathcal{M}}(\tau)$ belongs to $\overline{\Psi}(\tau)$ (such an optimal stationary policy exists according to Theorem 7.1.9 in [32]). Next we show that $\overline{\Psi}(\tau)$ is convex for every $\tau\in\Delta(S\times A)$. By the renewal theorem (Theorem 3.3.4 in [38]) the total reward occupation measure corresponding to some stationary policy f is equal to the average reward occupation measure under the same policy multiplied by the expected lifetime. This implies that for any $\rho\in\overline{\Psi}(\tau)$ (the notation used here is the same as in the definition of $\overline{B}(\tau)$):

$$\overline{G}(\tau)=\overline{J}^{\tau}(f^{\rho},Q(\cdot|s_0,a_0,\tau))=E^{Q,f}\mathcal{T}\sum_{s\in S}\sum_{a\in A}\rho_{sa}u(s,a,\tau).\tag{8}$$

Next note that the time spent in state s_0, playing action a_0 is independent of the policy used by the player. We shall denote this time by $\mathcal{T}_0$. Again by the renewal theorem we can write

$$E^{Q}\mathcal{T}_0=E^{Q,f}\mathcal{T}\frac{\rho_{s_0a_0}}{\sum_{s\in S}\sum_{a\in A}\rho_{sa}}=E^{Q,f}\mathcal{T}\rho_{s_0a_0}.$$

Substituting this into (8) we obtain

$$\overline{G}(\tau)=\frac{E^{Q}\mathcal{T}_0}{\rho_{s_0a_0}}\sum_{s\in S}\sum_{a\in A}\rho_{sa}u(s,a,\tau)$$

or equivalently

$$E^Q \mathcal{T}_0 \sum_{s\in S}\sum_{a\in A} \rho_{sa} u(s,a,\tau) - \overline{G}(\tau)\rho_{s_0 a_0} = 0.$$

The set of probability measures ρ satisfying the above equality is clearly a polytope, and hence a convex set. What we are left to show is that the graph of $\overline{\Psi}$ is closed. Suppose that $\tau_n, \tau, \rho_n, \rho \in \Delta(S\times A)$, $\tau_n \to \tau$, $\rho_n \to \rho$ and $\rho_n \in \overline{\Psi}(\tau_n)$ for every n. By Lemma 1 $\rho \in \overline{C}(\tau)$. Clearly, as $\rho_n \to \rho$, also $f^{\rho_n} \to f^{\rho}$. Since $\rho_n \in \overline{B}(\tau_n)$,

$$\overline{G}(\tau_n) = \overline{J}^{\tau_n}(f^{\rho_n}, Q(\cdot|s_o,a_0,\tau_n)) \geq \overline{J}^{\tau_n}(g, Q(\cdot|s_o,a_0,\tau_n))$$

for any stationary strategy g. By Lemma 3 and the continuity of Q also

$$\overline{J}^{\tau}(f^{\rho}, Q(\cdot|s_o,a_0,\tau)) \geq \overline{J}^{\tau}(g, Q(\cdot|s_o,a_0,\tau)),$$

implying that $\rho \in \overline{B}(\tau)$ and hence also in $\overline{\Psi}(\tau)$. This means that all the assumptions of the Glicksberg theorem [36] are satisfied and $\overline{\Psi}$ has a fixed point.

Now suppose τ^* is this fixed point. Because $\tau^* \in \overline{C}(\tau^*)$, it satisfies

$$\sum_{a\in A} \tau^*_{sa} = \sum_{x\in S}\sum_{b\in B} Q(s|x,b,\tau^*)\tau^*_{xb} = \Phi(s|\tau^*).$$

On the other hand, notice that

$$\overline{J}(\tau^*_S, g, f^{\tau^*}) = \overline{J}^{\tau^*}(g,\rho)$$

for any stationary policy g and any initial state distribution ρ, and so $\tau^* \in \overline{B}(\tau^*)$ implies that f^{τ^*} is the best response of a player when the state-action distribution of his oponents is always τ^*, and hence (f^*, τ^*_S) with $f^*(s,\tau) \equiv f^{\tau^*}(s)$ for any $\tau \in \Delta(S\times A)$ is an equilibrium in the game. □

5 The Relation with Games with Finitely Many Players

Main point of criticism of anonymous games in general is that the limiting situation with an infinite number of players does not exist in reality, and thus it is not sure that the results obtained for a continuum of players are relevant for the real-life ones, when the number of players is finite, but large. In this section we present some results connecting the anonymous game models from previous sections with similar models with finitely many players. In these results we will, in addition, use the following two assumptions.

(AT1) $Q(\cdot|a,s,\tau) = Q(\cdot|a,s)$ for all $\tau \in \Delta(S\times A)$ and $A(\cdot,\mu) = A(\cdot)$ for all $\mu \in \Delta(S)$.

(AT2) For any $f \in \mathcal{U}$ and $\tau \in \Delta(S\times A)$ the Markov chain of individual states of an individual using f when the state-action distribution of all the other players is τ is aperiodic.

The assumptions similiar to (AT1) appear also in a recent paper [39] on stochastic games with a finite number of players and average reward. They are also used in some papers on Markov evolutionary games [29,30] and in a recent application of anonymous sequential games to model power control problem in a wireless network [27]. It allows to decouple the Markov chains of individual states for each of the players (so that the dependence between different players is only through rewards) – this decoupling is crucial in our proofs of convergence of games with finite number of players to respective anonymous models. Importantly though, in most engineering applications individual state of a player is either his energy (or some other private resource) level or his geographical position. In both cases assumption (AT1) is naturally satisfied.

We will need to define some additional notation.

- We say that an n-person stochastic game is the n-person counterpart of an anonymous game if it is defined with the same objects S, A, u and Q, with the difference that the number of players is n and in consequence the global states and state-action distributions are defined on subsets of $\Delta(S)$ and $\Delta(S \times A)$,

$$\Delta^n(S) := \{\mu \in \Delta(S) : \mu_s = \frac{k_s}{n}, k_s \in \mathbb{N}, \text{ for all } s \in S\},$$

$$\Delta^n(S \times A) := \{\tau \in \Delta(S \times A) : \sum_{a \in A} \tau_{sa} = \frac{k_s}{n}, k_s \in \mathbb{N}, \text{ for all } s \in S\}.$$

- We will consider a wider set of possible policies in the game. Namely, we will consider a situation when each player uses a stationary policy over the whole game, but this policy is chosen at the beginning of the play according to some probability distribution. This means that any probability distribution from $\Delta(\mathcal{U})$ will be a policy in the game.
- We will consider a different (i.e. standard) definition of policies and equilibrium in the average-reward game. We will say that policies $(f^1, \ldots, f^n)$ form a Nash equilibrium in the average-reward n-person game if for any player i and any initial distribution of the global state μ^1,

$$J^i(\mu^1, f^1, \ldots, f^i, \ldots, f^n) \geq J^i(\mu^1, f^1, \ldots, g^i, \ldots, f^n)$$

for any other policy g^i. If this inequality holds up to some ε, we say that $(f^1, \ldots, f^n)$ are in ε-equilibrium. For both models (with average and total reward) we will also consider the notion of equilibrium defined as for the anonymous game – we will call it then a *weak equilibrium* (and analogously define weak ε-equilibrium).

Now we can prove the following two results:

Theorem 3 *Suppose (f, μ) is an equilibrium in either an average reward anonymous game satisfying (A1) and (AT1) or a total reward anonymous game satisfying (T1) and (AT1). Then for every $\varepsilon > 0$ there exists an $\overline{n}_\varepsilon$ such that for every $n \geq \overline{n}_\varepsilon$ (f, μ) is a weak equilibrium in the n-person counterpart of this anonymous game.*

A stronger result is true for the average reward game:

Theorem 4 *For every $\varepsilon > 0$ there exists an n_ε such that for every $n \geq n_\varepsilon$ the n-person counterpart of the average-reward anonymous game satisfying (A1), (AT1) and (AT2) has a symmetric Nash equilibrium $(\pi^n, \ldots, \pi^n)$, where $\pi^n \in \Delta(\mathcal{U})$. Moreover if (f, μ) is an equilibrium in the anonymous game, then π^n is of the form:*

$$\pi^n(s) = \sum_l \mu_l^* \delta[f_l^n(s)],$$

where $f_l^n(s) = \begin{cases} \overline{f}(s) & \text{if } s \notin S_l \\ f(s) & \text{if } s \in S_l \end{cases}$, $\overline{f}$ is the communicating policy induced by the assumption (A1)[8], S_l are ergodic classes of the individual state process of a player when he applies policy f and μ^* is the probability measure on these ergodic classes corresponding to measure μ over S[9].

Proof of Theorem 3: Let $J^l(\mu, g, f)$ denote the reward in the n-person counterpart of the given average-reward anonymous game for a player using stationary strategy g against f of all the others when initial distribution of individual states is μ.

$$\begin{aligned} J^n(\mu, g, f) &= \lim_{T\to\infty} \frac{1}{T} E^{g,f,\mu,Q} \sum_{t=1}^{T} u(s_t, g(x_t), \tau^t) \\ &= E^{f,\mu,Q} \sum_{x,a} \sigma_{xa}(g) m_f^n(\tau) u(x, a, \tau), \end{aligned}$$

where $\sigma(g)$ is the occupation measure of the process of individual states of the player using policy g and m_f^n is a probability measure over the set $\Delta(S \times A)$ denoting the frequency of the appearance of different state-action measures of all the players over the course of the game. Note further that $\sigma(g)$ is under (AT1) independent of n and the same as in the limiting (anonymous game) case. On the other hand m_f^n converges weakly as n goes to infinity to the invariant measure corresponding to the policy f and the initial distribution μ. Thus

$$J^n(\mu, g, f) \to_{n\to\infty} J(\mu, g, f)$$

for any stationary strategy g. The thesis of the theorem follows immediately.

To prove the total reward case we first need to write the reward in the n-person counterpart of the given total reward anonymous game for a player using stationary strategy g against f of all the others when initial distribution of states is μ.

$$\overline{J}^n(\mu, g, f) = E^{g,f,\mu,Q} \sum_{t=1}^{\mathcal{T}} u(s_t, g(x_t), \tau^t).$$

[8] Here it does not depend on τ by (AT1).

[9] Note that a measure over the set of states of an aperiodic Markov process can be invariant only if it is a convex combination of unique invariant measures on each of the ergodic classes. The coefficients of this convex combination are exactly the μ_l^*.

with s_1 distributed according to $Q(\cdot|s_0, a_0)$ (recall that by (AT1) this distribution is independent of the global state-action distribution τ). This by the renewal theorem equals

$$E^{g,\mu,Q}\mathcal{T}(g)\sum_{x,a}\sigma_{xa}(g)m_f^n(\tau)u(x,a,\tau),$$

where $\sigma(g)$ is the occupation measure of the process of individual states of the player using policy g and m_f^n is a probability measure over the set $\Delta(S\times A)$ denoting the frequency of the appearance of different state-action measures of all the players over the course of the game. But both $\sigma(g)$ and $E^{g,\mu,Q}\mathcal{T}(g)$ are under (AT1) independent of n and the same as in the limiting (anonymous game) case, while m_f^n converges weakly to the invariant measure corresponding to the policy f and the initial distribution μ as $n\to\infty$. The thesis of the theorem is now obtained as in the average reward case. □

Proof of Theorem 4: Fix $\varepsilon>0$ and take n_ε such that for every $n\geq n_\varepsilon$ $(f,\ldots,f)$ is a weak ε-equilibrium in the n-person counterpart of the average-reward game. Next note that for every $n\in\mathbb{N}$ the process of individual states of each of the players using policy π^n has a unique invariant measure μ. Since by (AT2) the process of individual states is aperiodic, it will converge to this invariant measure and, consequently, the reward for a player using policy f against f of all the others will be equal to $J(\mu,f,f)$ for any initial state distribution. However, since in a weakly communicating Markov decision process the optimal gain is independent of this distribution and since $(f,\ldots,f)$ is a weak ε-equilibrium in the n-person game for $n\geq n_\varepsilon$, it is also a Nash ε-equilibrium in the n-person game. □

6 Application: Medium Access Game

In the remainder of the paper we present two simple examples of application of our framework to model some real-life phenomena.

6.1 The Model

Consider the following MAC (Medium ACcess) game between mobile phones. Time is slotted. At any given time t, a mobile finds itself competing with N_t other mobiles for the access to a channel. N_t is assumed to have Poisson distribution with parameter λ. We shall formulate this as a sequential anonymous game as follows.

- **Individual state** A mobile has three possible states: F (full) AE (Almost Empty) and E (Empty).
- **Actions** There are two actions: transmit at high power H or low power L. At state AE a mobile cannot transmit at high power, while at E it cannot transmit at all.

- **Transition probabilities** From state AE the mobile moves to state E with probability p_E and otherwise remains in AE. At state E the mobile has to recharge. It moves to state F after one time unit. A mobile in state F transmitting with power r moves to state AE with probability proportional to r and given by αr for some constant $\alpha > 0$.
- **Payoff** Consider a given cellular phone that transmits a packet. Assume that x other packets are transmitted with high power and y with low power to the same base station. A packet transmitted with low power is received successfully with some probability q if it is the only packet transmitted, i.e. $y = 0$, $x = 0$. Otherwise it is lost. A packet transmitted with high power is received successfully with some probability $Q > q$ if it is the only packet transmitted at high power, i.e. $x = 0$. The immediate payoff is 1 if the packet is successfully transmitted. It is otherwise zero. In addition there is a constant cost $c > 0$ for recharging the battery. Aggregate utility for a player is then computed as long-time average of the per-period payoffs.

Suppose p is the fraction of population that transmits at high power in state F, and that μ_F, μ_{AE} and μ_E are fractions of players in respective states. Then probability of success for a player transmitting at high power is

$$QP(x=0) = Q(e^{-\lambda} + \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda}(1-p\mu_F)^k = Qe^{-\lambda}e^{\lambda(1-p\mu_F)} = Qe^{-\lambda p\mu_F},$$

while probability of success when a player transmits at low power is

$$qP(x+y=0) = q(e^{-\lambda} + \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda}\mu_E^k) = qe^{-\lambda}e^{\lambda\mu_E} = qe^{\lambda(\mu_E-1)}.$$

These values do not depend on actual numbers of players applying respective strategies – only on fractions of players in each of the states using different actions. Thus instead of considering an n-player game for any fixed n it is reasonable to apply the anonymous game formulation with $\tau = [\tau_{F,H}, \tau_{F,L}, \tau_{AE,L}, \tau_E]$ denoting the vector of fractions of players in respective states and using respective actions, with immediate rewards

$$u(s,a,\tau) = \begin{cases} Qe^{-\lambda\tau_{F,H}}, & \text{when } a = H \\ qe^{\lambda(\tau_E-1)}, & \text{when } a = L \\ -c, & \text{when } s = E \end{cases}$$

and transition probabilities defined by matrix

$$\mathbb{Q}(a,\tau) = \begin{bmatrix} 1-\alpha a & \alpha a & 0 \\ 0 & 1-p_E & p_E \\ 1 & 0 & 0 \end{bmatrix}.$$

6.2 The Solution

The stationary state of the chain of the private states of a player using policy f prescribing him to use high power with probability p when in state F is

$$\frac{1}{\alpha(pH+(1-p)L)(p_E+1)+p_E}\left[p_E, \alpha(pH+(1-p)L), p_E\alpha(pH+(1-p)L)\right].$$

Thus computations yield that his respected long-run average reward is of the form

$$\frac{Ap+B}{Cp+D} \quad \text{with}$$
$$A = p_E Q e^{-\lambda\tau_{F,H}} + ((H-L)\alpha - p_E)qe^{\lambda(\tau_E-1)} - c\alpha p_E(H-L),$$
$$B = (L\alpha+p_E)qe^{\lambda(\tau_E-1)} - c\alpha p_E L,$$
$$C = \alpha(H-L)(p_E+1),$$
$$D = \alpha L(p_e+1)+p_E.$$

It can be either a strictly increasing, a constant or a strictly decreasing function of p, depending on whether $AD > BC$, $AD = BC$ or $AD < BC$, and thus the best response of a player against the aggregated state-action vector τ is $p = 1$ when $AD > BC$, any $p \in [0,1]$ when $AD = BC$ or $p = 0$ when $AD < BC$. This leads to the following conclusion: since by Theorem 1 this anonymous game has an equilibrium, one of the three following cases must hold:

(a) If

$$\left[p_E Q e^{-\frac{\lambda p_E}{\alpha H(p_E+1)+p_E}} + ((H-L)\alpha - p_E)qe^{-\frac{\lambda(\alpha H+p_E)}{\alpha H(p_E+1)+p_E}} - a\alpha p_E(H-L)\right]$$
$$[\alpha L(p_E+1)+p_E] > \left[(L\alpha+p_E)qe^{-\frac{\lambda(\alpha H+p_E)}{\alpha H(p_E+1)+p_E}} - c\alpha L p_E\right][\alpha(H-L)(p_E+1)]$$

then all the players use high power in state F at equilibrium.

(b) If

$$\left[p_E Q e^{-\frac{\lambda p_E}{\alpha L(p_E+1)+p_E}} + ((H-L)\alpha - p_E)qe^{-\frac{\lambda(\alpha L+p_E)}{\alpha L(p_E+1)+p_E}} - a\alpha p_E(H-L)\right]$$
$$[\alpha L(p_E+1)+p_E] < \left[(L\alpha+p_E)qe^{-\frac{\lambda(\alpha L+p_E)}{\alpha L(p_E+1)+p_E}} - c\alpha L p_E\right][\alpha(H-L)(p_E+1)]$$

then all the players use low power in state F at equilibrium.

(c) If none of the above inequalities holds than we need to find p^* satisfying

$$\left[p_E Q e^{-\lambda\tau_{F,H}} + ((H-L)\alpha - p_E)qe^{\lambda(\tau_E-1)} - c\alpha p_E(H-L)\right][\alpha L(p_e+1)+p_E]$$
$$= \left[(L\alpha+p_E)qe^{\lambda(\tau_E-1)} - c\alpha p_E L\right][\alpha(H-L)(p_E+1)]$$

with $\tau_{F,H} = \frac{p^* p_E}{\alpha(p^*H+(1-p^*)L)(p_E+1)+p_E}$ and $\tau_E = \frac{\alpha(p^*H+(1-p^*)L)p_E}{\alpha(p^*H+(1-p^*)L)(p_E+1)+p_E}$. Then all the players use policy prescribing to use high power with probability p^* in state F at equilibrium.

Remark 3 It is worth noting here that some generalizations of the model presented above can be considered. We can assume that there are more energy levels and more powers at which players could transmit in our game (similiarly as in [27]). We can also assume that the players do not always transmit, only with some positive probability (then the individual state becomes two-dimensional, consisting of player's energy state and an indicator of whether he has something to transmit or not). Both these generalizations are tractable within our framework, though the computations become more involved.

7 The Case of Linear Utility: Maintenance-Repair Example

7.1 General Linear Framework

In the next example we consider a game satisfying some additional assumptions. Let $K = (S \times A)$. Let $\mathbf{u}(\tau)$ be a column vector whose entries are $u(k, \tau)$. We consider now the special case that $u(k, \tau)$ is linear in τ.

Equivalently, there are some vector $\mathbf{u^1}$ over K and a matrix $\mathbf{u^2}$ of dimension $|K| \times |K|$ such that

$$\mathbf{u}(\tau) = \mathbf{u}^1 + \mathbf{u}^2\tau$$

Similarly, we assume that the transition probabilities are linear in τ. Then the game becomes equivalent to solving a symmetric bilinear game, that of finding a fixed point of (2)-(3). Linear complementarity formulation can be used and solved using Lemke's algorithm. From the solution τ, the equilibrium (f, τ_S) can be derived with a help of equation (4).

7.2 Maintenance-Repair Game: The Model

The maintenence-repair example presented below can be seen as a toy model. Its main purpose though is to show how the abovementioned method can be used in a concrete game satisfying the linearity conditions mentioned above.

Each car among a large number of cars is supposed to drive one unit of distance per day. A car is in one of the **individual states** good (g) and bad (b). When a car is in a bad state then it has to go through some maintenance and repair actions and cannot drive for some (geometrically distributed) time.

A single driver is assumed to be infinitesimaly "small" in the sense that its contribution to the congestion experienced by other cars is negligible.

We assume that there are two types of behavior of drivers. Those that drive gently, and those that take risks and drive fast. This choice is modeled mathematically through **two actions:** aggressive (α) and gentle (γ). An aggressive driver is assumed to drive β times faster than a gentle driver.

Utilities A car that goes β time faster than another car, traverses the unit of distance at a time that is β times shorter. Thus the average daily delay it experiences is β times shorter. We assume that at a day during which a car drives fast, it spends $1/\beta$ of the time that the others do. It is then reasonable

to assume that the contribution to the total congestion is β times lower than that of the other drivers. More formally, let η be a delay function. Then the daily congestion cost D of a driver is given as

$$u(g, \alpha, \tau) = u(g, \gamma, \tau)/\beta$$

$$u(g, \gamma, \tau) = -\eta\left(\tau(g, \gamma) + \tau(g, \alpha)/\beta\right))$$

For the state b we set simply

$$u(b, a, \tau) = -1$$

which represents a penalty for being in a non-operational state. It does not depend on a nor τ.

Transition probabilities: We assume that transitions from g to b occur due to collisions between cars. Further assume that the collision intensity between a car that drives at state g and uses action a are linear in τ. More precisely,

$$Q(b|g, a, \tau) = c_a^\gamma \tau(g, \gamma) + c_a^\alpha \tau(g, \alpha).$$

We naturally assume that $c_\alpha^a > c_\gamma^a$ for $a = \alpha, \gamma$ and that $c_a^\alpha > c_a^\gamma$ for $a = \gamma, \alpha$. If a driver is more aggressive than another one, or if the rest of the population is more aggressive then the probability of a transition from g to b increases. We rewrite the above as

$$Q(b|g, a, \tau) = c_a \cdot \tau(g, .)$$

If a randomized stationary policy is used which chooses (α, γ) with respective probabilities $(p_\alpha, p_\gamma) =: \mathbf{p}$ then the one step transition from Q to b occurs with probability

$$Q(b|g, \mathbf{p}, \tau) = \sum_{a=\gamma,\alpha} p_a c_a \cdot \tau(g, .) =: \mathbf{p} \cdot c \cdot \tau(g, .).$$

Once in state b, the time to get fixed does not depend any more on the environment, and the drivers do not take any action at that state. Thus $\psi := Q(g|b, a, \tau)$ is some constant that is the same for all a and τ.

7.3 The Solution

We shall assume throughout that the congestion function η is linear. It then follows that this problem falls into the category of Section 7.1.

Let τ be given. Let a driver use a stationary policy $\mathbf{p}$. Then the expected time it remains in state g is

$$\sigma(\mathbf{p}, \tau) = \frac{1}{Q(b|g, \mathbf{p}, \tau)}$$

Its total expected utility during that time is

$$W_g(\mathbf{p},\tau) = \sigma(\mathbf{p},\tau)\sum_a p_a u(g,a,\tau) = \sigma(\mathbf{p},\tau)\sum_a p_a u(g,a,\tau)$$

$$= -\sigma(\mathbf{p},\tau)\left(p_\gamma + \frac{p_\alpha}{\beta}\right)\eta\left(\tau(g,\gamma) + \tau(g,\alpha)/\beta\right))$$

The expected repair time of a car (the period that consists of consecutive time it is in state b) is given by ψ^{-1}. Thus the total expected utility during that time is

$$W_b(\mathbf{p},\tau) = -\psi^{-1}.$$

Thus the average utility is given by

$$J(\mu,\mathbf{p},\pi(\tau)) = \frac{W_g(\mathbf{p},\tau)\psi - 1}{\frac{\psi}{Q(b|g,\mathbf{p},\tau)} + 1}$$

where μ is an arbitrary initial distribution and where $\pi(\tau)$ is the stationary policy that is obtained from τ as in (4).

Let $\mathbf{p}^*$ be a stationary equilibrium policy and assume that it is not on the boundary, i.e. $0 < p^*_\alpha < 1$. We shall consider the equivalent bilinear game. Let ρ^* be the occupation measure corresponding to $\mathbf{p}^*$. It is an equilibrium in the bilinear game.

Since the objective function is linear in ρ, ρ^* should be such that each individual player is indifferent between any stationary policy. In particular, we should have $J(\mu, 1_\alpha, \pi(\tau)) = J(\mu, 1_\gamma, \pi(\tau))$ where 1_a is the stationary pure policy that chooses always a.

We thus obtain the equilibrium occupation measure ρ^* as a τ that satisfies:

$$\frac{W_g(1_\alpha,\tau)\psi - 1}{\frac{\psi}{Q(b|g,\alpha,\tau)} + 1} = \frac{W_g(1_\gamma,\tau)\psi - 1}{\frac{\psi}{Q(b|g,\gamma,\tau)} + 1}.$$

The equilibrium policy $\mathbf{p}^*$ is obtained from ρ^* as in (4) and the equilibrium stationary measure is ρ^*_S.

8 Discussion and Conclusions

The framework of the game that is defined in this paper is similar in nature to the classical traffic assignment problem in that it has an infinity of players. In both frameworks, players can be in different states. In the classical traffic assignment problem, a class can be characterized by a source-destination pair, or by a vehicle type (car, pedestrian or bicycle). In contrast to the traffic assignment problem, the class of a player in our setting can change in time. Transition probabilities that govern this change may depend not only on the individual's state, but also on the fraction of players that are in each individual state and that use different actions. Furthermore, these transitions are controlled by the player.

A strategy of a player of a given class in the classical traffic assignment problem can be identified as the probability it would choose a given action (path) among those available to its class (or its "state"). The definition of a strategy in our case is similar, except that now the probability for choosing different actions should be specified not just in one state.

9 Acknowledgements

We would like to thank two anonymous referees for suggesting some important changes which helped us to improve the manuscript and Andrzej S. Nowak for some literature suggestions. The work of the second author has been partially supported by the European Commission within the framework of the CONGAS project FP7-ICT-2011-8-317672.

References

1. Wardrop, J.G.: Some theoretical aspects of road traffic research. Proc. Inst. Civ. Eng. 2, 325–378 (1952)
2. Haurie, A., Marcotte, P.: On the relationship between Nash-Cournot and Wardrop equilibria. Networks 15(3), 295–308 (1985)
3. Borkar, V.S.: Cooperative dynamics and Wardrop equilibria. Systems Control Letters 58(2), 91–93 (2009)
4. Maynard Smith, J.: Game Theory and the Evolution of Fighting. In: Maynard Smith, J. (ed.): On Evolution, pp. 8–28. Edinburgh University Press (1972)
5. Cressman, R.: Evolutionary Dynamics and Extensive Form Games. MIT Press, Cambridge, MA (2003)
6. Vincent, T.I., Brown, J.S.: Evolutionary Game Theory, Natural Selection and Darwinian Dynamics. Cambridge University Press, New York (2005)
7. Jovanovic, B., Rosenthal, R.W.: Anonymous Sequential Games. Journal of Mathematical Economics 17, 77–87 (1988)
8. Bergin, J., Bernhardt, D.: Anonymous sequential games with aggregate uncertainty. Journal of Mathematical Economics 21, 543–562 (1992)
9. Bergin, J., Bernhardt, D.: Anonymous Sequential Games: Existence and Characterization of Equilibria. Economic Theory 5(3), 461–89 (1995)
10. Sleet, C.: Markov perfect equilibria in industries with complementarities. Economic Theory 17(2), 371–397 (2001)
11. Chakrabarti, S.K.: Pure strategy Markov equilibrium in stochastic games with a continuum of players. J. Math. Econom. 39(7), 693ï£¡-724 (2003)
12. Adlakha, S., Johari, R.: Mean Field Equilibrium in Dynamic Games with Strategic Complementarities. Operations Research, 61(4), 971–989 (2013)
13. Green, E.: Noncooperative Price Taking in Large Dynamic Markets. Journal of Economic Theory 22, 155–181 (1980)
14. Green, E.: Continuum and Finite–Player Noncooperative Models of Competition. Econometrica 52, 975–993 (1984)
15. Housman, D.: Infinite Player Noncooperative Games and the Continuity of the Nash Equilibrium Correspondence. Mathematics of Operations Research 13, 488–496 (1988)
16. Sabourian, H.: Anonymous repeated games with a large number of players and random outcomes. Journal of Economic Theory 51(1), 92–110 (1990)
17. Krusell, P., Smith, A.A. Jr.: Income and Wealth Heterogeneity in the Macroeconomy. Journal of Political Economy 106, 867ï£¡-896 (1998)
18. Hopenhayn, H.A.: Entry, exit, and firm dynamics in long run equilibrium. Econometrica 60(5), 1127-1150 (1992)

19. Hopenhayn, H.A., Prescott, E.C.: Stochastic monotonicity and stationary distributions for dynamic economies. Econometrica 60(6), 1387-ï£¡1406 (1992)
20. Weintraub, G.Y., Benkard, C.L., Van Roy, B.: Markov Perfect Industry Dynamics with Many Firms. Econometrica 76, 1375–1411 (2008)
21. Weintraub, G.Y., Benkard, C.L., Van Roy, B.: Industry dynamics: Foundations for models with an infinite number of firms. J. Econom. Theory 146(5), 1965ï£¡-1994 (2011)
22. Krishnamurthy, I., Johari, R., Sundararajan, M.: Mean field equilibria of dynamic auctions with learning. ACM SIGecom Exchanges (2011)
23. Bodoh–Creed, A.: Approximation of Large Games with Application to Uniform Price Auctions. forthcoming in Journal of Economic Theory (2012)
24. Bodoh–Creed, A.: Optimal Platform Fees for Large Dynamic Auction Markets. mimeo (2012)
25. Karatzas, I., Shubik, M., Sudderth, W.D.: Construction of Stationary Markov Equilibria in a Strategic Market Game. Math. Oper. Res. 19(4), 975–1006 (1992)
26. Karatzas, I., Shubik, M., Sudderth, W.D.: A Strategic Market with Secured Lending. J. Math. Econ. 28, 207–247 (1997)
27. Więcek, P., Altman, E., Hayel, Y.: Stochastic State Dependent Population Games in Wireless Communication. IEEE Transactions on Automatic Control 56(3), 492–505 (2011)
28. Tembine, H., Lasaulce, S., Jungers, M.: Joint power control-allocation for green cognitive wireless networks using mean field theory. IEEE Proc. of the 5th Intl. Conf. on Cogntitive Radio Oriented Wireless Networks and Communications, (2010)
29. Altman, E., Hayel, Y.: Stochastic Evolutionary Games. Proceedings of the 13th Symposium on Dynamic Games and Applications, Wroclaw, Poland, 30th June–3rd July, (2008)
30. Altman, E., Hayel, Y., Tembine, H., El-Azouzi, R.: Markov decision Evolutionay Games with Time Average Expected Fitness Criterion. 3rd International Conference on Performance Evaluation Methodologies and Tools, (Valuetools), Athens, Greece, 21–23 October, (2008)
31. Ross, K.W., Varadarajan, R.: Multichain Markov Decision Processes with a sample path constraint: A decomposition approach. Mathematics of Operations Research 16(1), 195–207 (1991)
32. Puterman, M.: Markov Decision Processes. Wiley-Interscience, New York (1994)
33. Bather, J.: Optimal Decision Procedures in Finite Markov Chains, Part II: Communicating Systems, Adv. in Appl. Probability 5, 521–552 (1973)
34. Bather, J.: Optimal Decision Procedures in Finite Markov Chains, Part I: Examples. Adv. in Appl. Probability 5, 328–339 (1973)
35. Mannor, S., Tsitsiklis, J.N.: On the Empirical State-Action Frequencies in Markov Decision Processes Under General Policies. Mathematics of Operations Research 30(3), 545–561 (2005)
36. Glicksberg, I.L.: A further generalization of the Kakutani fixed point theorem with application to Nash equilibrium points. Proc. Amer. Math. Soc. 3, 170–174 (1952)
37. Altman, E.: Constrained Markov Decision Processes. Chapman & Hall (1999)
38. Ross, S.M.: Stochastic Processes. 2nd Edition, John Wiley & Sons (1996)
39. Flesch, J., Schoenmakers, G., Vrieze, K.: Stochastic games on a product state space: the periodic case. Int. J. Game Theory 38, 263–289 (2009)