Stationary anonymous sequential games with undiscounted rewards

Piotr Wiecek, Eitan Altman

Stationary anonymous sequential games with undiscounted rewards

Piotr Więcek
Institute of Mathematics and Computer Science
Wrocław University of Technology
Wybrzeże Wyspiańskiego 27
50-370 Wrocław, Poland
Email: Piotr.Wiecek@pwr.wroc.pl

Eitan Altman
INRIA
2004 Route des Lucioles, P.B. 93
06902 Sophia Antipolis Cedex, France
Email: Eitan.Altman@sophia.inria.fr

***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 cost as well as to the expected average cost. We show that equilibria in the anonymous sequential game correspond to the limit of equilibria of related finite population games as the number of players grow to infinity. We provide many examples to illustrate our results.**

I. Introduction

Games with a continuum set 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 [15] in 1952 to model the choice of routes of cars where each driver, modeled as an atomless player, minimizes its expected travel delay. In Wardropt'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 class. The equilibrium is known to behave as the limit of the equilibrium obtained in a game with finitely many players, as their number tends to infinity [8]. It is also the limit of some dynamic games in which randomness tends to average away as the number of players increase [4].

Another class of games that involves a continuum of atomless player is the evolutionary games, in which pairs of players that play a matrix game are selected at random, see [11]. The 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 a Evolutionary Stationary Strategy (we refer the reader to [5], [14]).

In both games described above, the playert's type is fixed, and the action 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 infinitely 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 sequential anonymous games, introduced by B. Jovanovic and R.W. Rosenthal in 1988 in [9]. In that work, each playert's utility is given as the expected discounted utility over an infinite horizon. Our contribution in this paper is to solve the cases of expected average utility and of the total expected utility which have remained open ever since 1988.

Similar extensions have been proposed and studied for the framework of evolutionary games in [1], [2]. The analysis there turns out to be simpler since the utility in each encounter between two players turns out to be bilinear there. In [16] we have already used the framework of anonymous sequential games to study a power control problem. Our work was restricted to the total expected cost criteria where as in this paper we focus in particular on the expected average cost, which, to the best of our knowledge, has not been stuied before.

The structure of the paper is as follows. Section II presents the model and introduces in particular the expected average and the total expected cost criteria. Several theoretical are stated without proof at the end of the section. We then study in detail a stochastic maintenantce game (Section III) which is followed by a concluding paragraph.

II. 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 a 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 would write $Q(\cdot|s_t, a_t, \tau^t)$ for the distribution of the individual state at time $t+1$, given his state s_t, his action a_t and the state-action distribution of all the players.
- The distribution of the global state at time $t + 1$ will be given by $\Phi(\cdot|\tau^t) = \sum_{s\in S}\sum_{a\in A} Q(\cdot|s, a, \tau^t)\tau^t_{sa}$.

any function $f : S \to \Delta(A)$ is called a *stationary policy*. If for a given measure μ on S it satisfies $\mathrm{supp} f(s) \subset A(s, \mu)$ for every $s \in S$, we call it *μ-feasible*. We denote the set of stationary policies in our game by $\mathcal{U}$, and the set of μ feasible stationary policies by $\mathcal{U}^\mu$.

A. 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 $f \in \mathcal{U}^\mu$, for every other stationary strategy $g \in \mathcal{U}^\mu$,

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

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

[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.

Remark 1: The definition of the equilibrium used here differs significantly from that used in [9]. There the equilibrium is defined with respect to the solution of some dynamic programming. Our definition directly relates it to the cost functionals.

B. 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 *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, 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 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 would be slightly different from that for the average cost. We define a stationary strategy f and a measure $\mu \in \Delta(S)$ to be in equilibrium in the total reward game if $f \in \mathcal{U}^\mu$, for every other stationary strategy $g \in \mathcal{U}^\mu$,

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

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 $t \geq 1$.

Remark 2: Note that although we assume that our anonymous game is symmetric, we can easily introduce asymmetry in our model. Namely, if we divide S into a (finite) number of subsets S^i such that under any policy it is impossible to move from one S^i to another one, we can model asymmetric anonymous game with a finite number of "types" (corresponding to different sets of individual states S^i) of the players. All of the results proved in the next sections remain true for this asymmetric version of the model, although the proofs of some of them may be slightly more involved in that case.

C. Assumptions

We introduce below some assumptions that will be used later.

The following assumptions will be used in Section II-D:

(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.

(A2) 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.

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. [12], chapters 8 and 9.

Remark 3: A set S is called communicating if for any pair of states there exists a pure stationary policy and an integer r such that the probability of reaching the second state from the first one in r step is strictly positive. If $S1$ is strongly communicating then it is communicating [13]. The converse need not hold (indeed the set S_0 can be communicating too but it is not strongly communicating). However, due to the assumption that S_0 is transient under all stationary policies, S_1 is closed (in the sense that it is impossible to leave it), and is thus strongly communicating (see [13]). See also [3].

The following assumptions will be used in Section II-D:

(T1) There exists a $p_0 > 0$ such that for any fixed state-action measure τ and under any τ_S-feasible 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 $p_0(\tau)$.

(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 $\mu \in \Delta(S)$.

This kind of assumption appears also in a recent paper [6] on stochastic games with a finite number of players and average reward.

D. Existence of equilibrium

We briefly state some theoretical results whose proof is omitted.

Theorem 1: Every anonymous sequential game with total reward satisfying (T1) has a stationary equilibrium.

Theorem 2: Every anonymous sequential game with long-time average payoff satisfying (A1) and (A2) has a stationary equilibrium.

Our game is the mean field limit of some games with a finite number n of players. This is summarized in the following.

Theorem 3: Suppose (f, μ) is an equilibrium in either average reward anonymous game satisfying (A1), (A2) and (AT1) or 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.

The proofs of the above Theorems as well as further stronger results will be available at the home page of the second author.

E. Games with linear utility

Let $K = (S \times A)$. Let $\mathbf{u}(\tau)$ be a column vector whose entries are $u(k, \tau)$. We consider in this section theh 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. Linear complementarity formulation can be used and solved using Lemke's algorithm.

III. A MAINTENANCE-REPAIR EXAMPLE

A. The Model

Each car among a large nuber 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 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 behaviors 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 f be a delay function. Then the daily congestion cost D of a driver is given as

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

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

For the state b we set simply

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

which represents a penalty for being in an 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 rewtite 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.

B. Solution

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

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

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

Its total expected utility during that time is

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

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

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

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

Thus the average utility is given by

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

where μ is an arbitrary initial distribution and where π is the stationary policy that is obtained from τ by

$$(\pi i(s))_a = \begin{cases} \frac{\tau_{sa}}{\sum_{b \in A} \tau_{sa}} & \text{if } \sum_{b \in A} \tau_{sa} > 0 \\ \\ \delta[a_0] & \text{otherwise} \end{cases} \quad (1)$$

where $\delta[x]$ denotes a probability measure concentrated in x.

Let $\mathbf{p}^*$ be a stationary equilibrium 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 by finding τ that satisfies:

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

IV. 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 gouvern 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.

References

[1] E. Altman and Y. Hayel, "Stochastic Evolutionary Games", *Proceedings of the 13th Symposium on Dynamic Games and Applications*, Wroclaw, Poland, 30th June-3rd July, 2008.

[2] E. Altman, Y. Hayel, H. Tembine, R. El-Azouzi, "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.

[3] J. Bather, Optimal Decision Procedures in finite Markov chains, part II: Communicating Systems, Adv in Appl. Probability, Vol 5, pp. 521-552, 1973.

[4] V. Borkar and P. R. Kumar, "Dynamic Cesaro-Wardrop Equilibration in Networks," IEEE Transactions on Automatic Control, pp. 382-396, vol. 48, no. 3, March 2003

[5] R. Cressman, *Evolutionary Dynamics and Extensive Form Games*, MIT Press, Cambridge, MA, 2003.

[6] J. Flesch, G. Schoenmakers and K. Vrieze, *Stochastic games on a product state space: the periodic case*. Int. J. Game Theory 38, 263–289.

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

[8] A. Haurie and P. Marcotte; On the relationship between Nash -Cournot and Wardrop equilibria, Networks, vol. 15, pp. 295 - 308, 1985.

[9] B. Jovanovic and R.W. Rosenthal, 1988, *Anonymous Sequential Games*, Journal of Mathematical Economics 17 77–87.

[10] S. Mannor and J.N. Tsitsiklis, 2005, *On the Empirical State-Action Frequencies in Markov Decision Processes Under General Policies*. Mathematics of Operations Research, 30(3), 545-561.

[11] J. Maynard Smith, "Game Theory and the Evolution of Fighting", in *On Evolution*, John Maynard Smith (Edr) Edinburgh: Edinburgh University Press, pp 8-28, 1972.

[12] M. Puterman, 1994, *Markov Decision Processes*. Wiley-Interscience, New York.

[13] K.W. Ross and R. Varadarajan, Multichain Markov Decision Processes with a sample path constraint: A decomposition approach. *Mathematics of Operations Research*, Vol 16 No 1, pp 195-207, 1991.

[14] T.I. Vincent and J.S. Brown, *Evolutionary Game Theory, Natural Selection and Darwinian Dynamics*, Cambridge University Press.

[15] J.G. Wardrop, "Some theoretical aspects of road traffic research", *Proc. Inst. Civ. Eng.* Part 2, 325–378, 1952.

[16] P. Więcek, E. Altman and Y. Hayel, "An Anonymous Sequential Game Approach for Battery State Dependent Power Control", proceedings of NET-COOP, Eurandom, the Netherlands, November 2009.