Markov Decision Evolutionary Games with Time Average Expected Fitness Criterion

[Invited paper] *

Eitan Altman
INRIA, MAESTRO Group
2004 Route des Lucioles
F-06902,Sophia-Antipolis Cedex
Altman@sophia.inria.fr

Yezekael Hayel
LIA/CERI
University of Avignon
339, chemin des Meinajaries
Agroparc BP 1228
F-84911 AVIGNON Cedex
Yezekael.Hayel@univ-avignon.fr

Hamidou Tembine
LIA/CERI
University of Avignon
339, chemin des Meinajaries
Agroparc BP 1228
F-84911 AVIGNON Cedex
hamidou.tembine@univ-avignon.fr

Rachid ElAzouzi
LIA/CERI,
University of Avignon
339, chemin des Meinajaries
Agroparc BP 1228
F-84911 AVIGNON Cedex
Rachid.Elazouzi@univ-avignon.fr

ABSTRACT

We present a class of evolutionary games involving large populations that have many pairwise interactions between randomly selected players. The fitness of a player depends not only on the actions chosen in the interaction but also on the individual state of the players. Players stay permanently in the system and participate infinitely often in local interactions with other randomly selected players. The actions taken by a player determine not only the immediate fitness but also the transition probabilities to its next individual state. We define and characterize the Evolutionary Stable Strategies (ESS) for these games and propose a method to compute them.

Keywords

Evolutionary stable strategy, Markov games.

1. INTRODUCTION

Evolutionary games have been developed by J. Maynard Smith to model the evolution of population sizes as a result of competition between them that occurs through many local pairwise interactions, i.e. interactions between randomly chosen pairs of individuals. Central in evolutionary games is the concept of Evolutionary Stable Strategy, which is a distribution of (deterministic or mixed) actions such that if used, the population is immune against penetration of mutations. This notion is stronger than that of Nash equilibrium as ESS is robust against a deviation of a whole fraction of the population where as the Nash equilibrium is defined with respect to possible deviations of a single player. A second foundation of evolutionary games is the replicator dynamics that describes the dynamics of the sizes of the populations as a result of the fitness they receive in interactions. Meynard Smith formally introduced both, without needing an explicit modeling of stochastic features. We shall call this the deterministic evolutionary game.

Randomness is implicitly hinted in the requirement of robustness against mutations, and indeed the ESS is defined through robustness against any mutation. Random aspects can be explicitly included in the modeling of evolutionary games. We first note that since deterministic evolutionary games deal with large populations, they may provide an interpretation of the deterministic game as a limit smaller games that included randomness that has been averaged out by some strong law of large numbers. Such an interpretation can be found in [11].

Yet, other of randomness have been introduced into evolutionary games. Some authors have added small noise to the replicator dynamics in order to avoid the problem of having the dynamics stuck in some local minimum, see [15, 12, 13] and references therein. The ESS can then be replaced by other notions such as the the SSE [12].

In this paper we introduce another class of stochastic evolu-

*This work was partially supported by the INRIA ARC Program POPEYE.

tionary games, which we call "Markov Decision Evolutionary Games" (MDEG). There are again many local interactions among individuals belonging to large populations of players. Each individual stays permanently in the system; from time to time it move among different individual states, and interacts with other users. The actions of the player along with those with which it interacts determine not only the immediate fitness of the player but also the transition probabilities to the next state it will have. Each individual is thus faced with an MDP in which it maximizes the expected average cost criterion. Each individual knows only the state of its own MDP, and does not know the state of the other players it interacts with. The transition probabilities of a player's MDP are only controlled by that player. The local interactions between players can be viewed as a cost-coupled stochastic game [1, 2] which suggests the sufficiency of stationary strategies.

A simple application of an MDEG to mobile communications has been introduced in [6] for the case in which individuals have finite life time and the criterion that is maximized is the total expected fitness during the individual's life time. Mobile terminals transmit packets occasionally. Their destination occasionally may receive simultaneously a transmission from another terminal which results in a collision. It is assumed however that even when packets collide, one of the packets can be received correctly if transmitted at a higher power. The immediate fitness rewards successful transmissions and penalizes energy consumption. Each mobile decides at each slot what its power level will be. This decision is allowed to depend on the depletion level of the battery, which serves as the "individual state". The battery is considered to be either in the state "Full" (F) in which case there are two power levels available, or "Almost Empty" (AE) in which only the weak power level is available, or at the empty state E. Transmission at high power at state F results in a larger probability of moving to state AE. When at state E, the battery is replaced by a new one at some constant cost. We extend this model in Subsection 3.4, and adapt it to the average expected fitness criterion.

An interesting application of MDEG is the repeated game version of the well known Hawk and Dove game in which some of the features of MDEG are already present, [7, 8, 9].

2. REMINDER ON (STANDARD) EVOLUTIONARY GAMES (EG)

Consider a large population of players. Each individual needs occasionally to take some action. We focus on some (arbitrary) tagged individual. Occasionally, the action of some N (possibly random number of) other individuals interact with the action of that individual. We define by $J(p, q)$ the expected payoff for our tagged individual if it uses a strategy (also called policy) p when meeting another individual who adopts the strategy q. This payoff is called "fitness" and strategies with larger fitness are expected to propagate faster in a population. p and q belong to a set K of available strategies. In the standard framework for evolutionary games there are a finite number of so called "pure strategies", and a general strategy of an individual is a probability distribution over the pure strategies. An equivalent interpretation of strategies is obtained by assuming that individuals choose pure strategies and then the probability distribution represents the fraction of individuals in the population that choose each strategy. Note that J is linear in p and q.

Suppose that the whole population uses a strategy q and that a small fraction ϵ (called "mutations") adopts another strategy p. Evolutionary forces are expected to select against p if

$$J(q, \epsilon p + (1-\epsilon)q) > J(p, \epsilon p + (1-\epsilon)q) \quad (1)$$

DEFINITION 2.1. *A strategy q is said to be ESS if for every $p \neq q$ there exists some $\bar{\epsilon}_y > 0$ such that (1) holds for all $\epsilon \in (0, \bar{\epsilon}_y)$.*

In fact, we expect that if

$$\text{for all } p \neq q, \quad J(q, q) > J(p, q) \quad (2)$$

then the mutations fraction in the population will tend to decrease (as it has a lower reward, meaning a lower growth rate). q is then immune to mutations. If it does not but if still the following holds,

$$\text{for all } p \neq q, \quad J(q, q) = J(p, q) \text{ and } J(q, p) > J(p, p) \quad (3)$$

then a population using q are "weakly" immune against a mutation using p since if the mutant's population grows, then we shall frequently have individuals with strategy q competing with mutants; in such cases, the condition $J(q, p) > J(p, p)$ ensures that the growth rate of the original population exceeds that of the mutants. We shall need the following characterization:

THEOREM 2.1. *[3, Proposition 2.1] or [4, Theorem 6.4.1, page 63] A strategy q is ESS if and only if it satisfies (2) or (3).*

COROLLARY 2.1. *(2) is a sufficient condition for q to be an ESS. A necessary condition for it to be an ESS is*

$$\textit{for all } p \neq q, \quad J(q, q) \geq J(p, q) \quad (4)$$

The conditions on ESS can be related and interpreted in terms of a Nash equilibrium in a matrix game. The situation in which an individual, say player 1, is faced with a member of a population in which a fraction p chooses strategy A is then translated to playing the matrix game against a second player who uses mixed strategies (randomizes) with probabilities p and $1-p$, resp.

3. MODEL

We use a hierarchical description of the system composed of a model for the individual player and a global model for aggregating individual's behavior.

3.1 Model for Individual player

A player arrives is born at some random time t_0. It has a clock that is responsible to the times at which interactions with other players occur. These interactions occur according to a Poisson process with rate λ. Each time the timer

clicks, the player interacts with another randomly selected player. It receives some reward (fitness) that depends on the individual state of the players involved in the interaction and on their action at that instant. We associate with each player a Markov Decision Process (MDP) embedded at the instants of the clicks.

The parameters of the MDP are given by the tuple $\{\mathcal{S}, \mathcal{A}, Q\}$ where

- $\mathcal{S}$ is the set of possible individual states of the player
- $\mathcal{A}$ is the set of available actions. For each state s, a subset $\mathcal{A}_s$ of actions is available.
- Q is the set of transition probabilities; for each $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}_s$, $Q_{s'}(s,a)$ is the probability to move from state s to state s' taking action a. $\sum_{s' \in S} Q_{s'}(s,a)$ is allowed to be smaller than 1.

Define further

- The set of (behavioral) policies is $\mathcal{U}$. A general policy u is a sequence $u = (u_1, u_2, \ldots)$ where u_i is a distribution over action space $\mathcal{A}$ at time i. The dependence on time is a local one: it concerns only the individual's clock; a player is not assumed to use policies that make use of some global clocks.
- The subset $\mathcal{U}_S$ of stationary policies; a stationary policy u is a policy in which the probability to choose an action a depends only on the current state s; it is denoted by $u(a|s)$.
- The subset $\mathcal{U}_D \subset \mathcal{U}_S$ of pure or deterministic stationary policies $\mathcal{U}_D$. A policy of this type can be viewed as a function from the states to the actions.
- The set $\mathcal{U}_M$ of mixed strategies: A mixed strategy is identified with a probability γ over the set of pure stationary strategies. It can be considered as first choosing a pure stationary policy u with probability $\gamma(u)$ and then keeping choosing forever the actions according to u.

Occupation measure Often we encounter the notion of individual states in evolutionary games; but usually the population size at a particular state is fixed. In our case the choices of actions of an individual determine the fraction of time it would spend at each state. Hence the fraction of the whole population that will be at a given state may depend on the distribution of strategies in the population. In order to model this dependence we first need to introduce the expected average frequency $f^t_{\eta,u}(s)$ that an individual spends till time t at a given state s when it follows a strategy u and its initial state at time 1 is distributed according to a probability η over $\mathcal{S}$. More generally, we define $f^t_{\eta,u}(s,a)$ the expected average frequency till time t during which it is at state s and it chooses action a.

More precisely, define $p_t(\eta, u; s, a) = I\!P_{\eta,u}(S_t = s, A_s = a)$ the probability for a user to be in state s, at time t, using action a under policy u when the initial state has a probability distribution η. Further define $p_t(\eta, u; s) = \sum_a p_t(\eta, u; s, a)$. Define

$$f^t_{\eta,u}(s) = \frac{1}{t}\sum_{r=1}^{t} p_r(\eta, u; s), \qquad f^t_{\eta,u}(s,a) = \frac{1}{t}\sum_{r=1}^{t} p_r(\eta, u; s, a)$$

Denote $f^t_\eta(u) := \{f^t_{\eta,u}(s,a)\}$.

Define the Φ^u_η to be the set of all accumulation points of $f^t_\eta(u)$ as $t \to \infty$. Whenever Φ^u_η contains a single element, we shall denote it by $f_\eta(u)$.

We shall assume the following:

A1: Under any pure stationary, policy S_t is unichain: it is a Markov chain that has a single ergodic class of states.

3.2 Interactions and System model

We have a large population of individuals. As in standard evolutionary games, there are many pairwise interactions between randomly selected pairs.

Let $r(s, a, s', b)$ be the immediate reward that a player receives when it is at state s and it uses action a while interacting with a player who is in state s' that uses action b.

Denote by $\alpha(u) = \{\alpha(u; s, a)\}$ the system state: $\alpha(u; s, a)$ is the fraction of the population at individual state s and that use action a when all the population uses strategy u. We shall add the index t to indicate a possible dependence on some time.

Consider an arbitrary tagged player and let S_t and A_t be its state and action at time t (as measured on its individual clock). Then his expected immediate reward at that time when all other players use u is given by

$$R_t = \sum_{s,a} \alpha_t(u; s, a) r(S_t, A_t, s, a).$$

Assume now that a player arrives at the system at (local) time 1. The expected fitness when using a policy v is then

$$F_\eta(v, u) = \liminf_{t\to\infty} \frac{1}{t}\sum_{m=1}^{t} E_{\eta,v}[R_m].$$

When η is concentrated on state s we write with some abuse of notation $F_s(v, u) = F_\eta(v, u)$. We shall often omit the index η (in case it is taken to be fixed).

Introduce the following assumptions.

A2(U): When the whole population uses a policy $u \in U$, then at any time t which is either fixed or is an individual time of an arbitrary player, $\alpha_t(u)$ is independent of t and is given by

$$\alpha_t(u; s, a) = f_{\eta,u}(s, a) = \pi(s)u(a|s)$$

for all s, a where $f_{\eta,u}(s, a)$ is the single limit of $f^t_{\eta,u}(s, a)$ as $t \to \infty$.

A2: Assumption $A2(U)$ holds for $U = U_s$ and for $U = U_M$.

The validity of the Assumption depends on the way the infinite population model is obtained by scaling a large finite population model. This aspect is beyond the scope of this paper.

Denote the set of all policies for which Φ_η^u is a singleton by $\overline{U}^*$.

For $u \in U^*$, the following holds:

$$F(v,u) = \inf_{z \in \Phi_\eta^v} \sum_{s,a} z(s,a) \sum_{s',a'} f_{\eta,u}(s',a') r(s,a,s',a'). \quad (5)$$

Note that for any $u \in U_M$, and for any strategies v and w,

$$\Phi_\eta^v \subset \Phi_\eta^w \quad \text{implies } F(v,u) \geq F(w,u). \quad (6)$$

This, together with the fact that for any policy u and $z \in \Phi_\eta^u$ there exist a stationary policy $v \in U^*$ satisfying $f_\eta^v = z$, will motivate us to limit ourselves to policies in U^*.

When both u and v are in U^*, the global expected fitness simplifies to

$$F(v,u) = \sum_{t=1}^{\infty} E_{\eta,v} R_t = \sum_{s,a} f_{\eta,v}(s,a) \sum_{s',a'} f_{\eta,u}(s',a') r(s,a,s',a'). \quad (7)$$

Definition 3.1. *We shall say that two strategies u and v in U^* are equivalent if the corresponding occupation measures are equal: $f_\eta^v = f_\eta^u$. We shall write $u =_e u'$.*

Note that if u and u' are equivalent policies for a given player then for any $v \in U^*$ used by the rest of the population, the fitness under u and under u' are the same.

3.3 Defining Equilibrium and Weak ESS

With the expression (7) for the fitness, we observe that we are again in the framework of Section 2.

Definition 3.2. *(i) A strategy $u \in U^*$ is an equilibrium for the MDEG if and only if it feasible and satisfies*

$$F(u,u) \geq F(v,u). \quad (8)$$

(ii) A strategy $u \in U^*$ is a weak ESS (WESS) for the MDEG if and only if

- *it is an equilibrium, and*
- *for all $v \in U^*$ such that $v \neq_e u$ that satisfy $F(u,u) = F(v,u)$, the following holds: $F(u,v) > F(v,v)$.*

The fitness function (5) is bilinear in the occupation measures of the players that interact with each other. The set of occupation measures will be shown to be a polytope whose extreme points correspond to strategies in U_D. This will allow us to transform the MDEG to a standard EG.

We could use the following as an equivalent Definition of WESS for MDEG.

Theorem 3.1. *A strategy $u \in U^*$ is a WESS if and only if for every $v \in U^*$ with $v \neq_e u$, there exists some $\bar{\epsilon}_v > 0$ such that the following holds for all $\epsilon \in (0, \bar{\epsilon}_v)$:*

$$F(u, \epsilon u + (1-\epsilon)v) > F(u, \epsilon u + (1-\epsilon)v) \quad (9)$$

In (9) we use a convex combination of two policies. We delay the definition of this to the next section (see Remark 4.1).

3.4 Application to Energy Control in Wireless Networks

We next illustrate the MDEG setting with a problem that arises in dynamic power control in mobile networks. A special case of this framework (where a choice between several control actions exists in one state only) has been studied in [6] with, however, a total cost criterion.

Users participate in local competitions for the access to a shared medium in order to transmit their packets. An individual state of each mobile represents the energy level at the user's battery which, for simplicity, we assume to take finitely many values, denoted by $\mathcal{S} = \{0, \ldots, n\}$.

Each time the battery empties (which corresponds to reaching state 0), the mobile changes the battery to a new one (this corresponds to state n), and pay a cost C. We assume that each time a mobile reaches state zero, it remains there during a period whose expected duration is τ.

In each state $s \in \mathcal{S} \setminus \{0\}$, each mobile has two available actions h and l which correspond respectively to high power p_H and low power p_L. We consider an Aloha-type game where a mobile transmits a packet with success during a slot if:

- with probability p, the mobile is the only one to transmit during this slot,
- the mobile transmits with high power and the other transmitting mobile uses low power or is in state 0.

The reward function r depends on a mobile's state as well as on the transmission powers, that is, the action of the mobile as well as that of the one it interacts with. Then we have for $s \neq 0$:

$$r(s,a,s',a') = p + (1-p)\mathbb{1}_{(s'=0)} + (1-p)\mathbb{1}_{((a=h),\ (a'=l),\ (s'\neq 0))}.$$

For $s = 0$ we take $r(0,a,s',a') = C/\tau$.

For each state $s \in \mathcal{S} \setminus \{0\}$, the transition probability $Q_{s'}(s,a)$ may be non-zero (for both $a \in \{l,h\}$) only for $s' \in \{s, s-1\}$. Then, as the two possible transitions are to remain at the same energy level or move to the next lower one, we simplify the notation and use $Q(s,a)$ to denote the probability of remaining at energy level s using action a.

To model the fact that the mobiles stays in the average τ units at state 0 and then moves to state n we set the transition probabilities from state 0 to any state other than n and 0 to be zero; the probability to move to n is $1/\tau$ and that of remaining at 0 is $1 - 1/\tau$.

We have the following assumptions on the transition probabilities which are motivated by the application.

- For all state $s \in \mathcal{S} \setminus \{0\}$, we have $Q(s,h) < Q(s,l)$ because using less power induces higher probability to remain in the same energy level.
- For all state $s \in \mathcal{S} \setminus \{0\}$ and for both actions $a \in \{l,h\}$, we have $Q(s,a) > Q(s-1,a)$ because less battery energy the mobile has, less is the probability to remain at the same energy level.

4. COMPUTING EQUILIBRIA AND WESS

Define the set of occupation measures achieved by all (individual) policies in some subset $U' \subset U$ as

$$\mathcal{L}_\eta(U') = \bigcup_{u \in U'} \Phi_\eta^u.$$

Recall from eq (7) that the expected fitness of an individual depends on the strategy u of that individual only through $f_{\eta,u}$. We are therefore interested in the following characteristic of $\mathcal{L}_\eta(U)$ (see [14, 10]):

LEMMA 4.1. *$\mathcal{L}_\eta(U)$ equals to the set Q_η defined as the set of $\alpha = \{\alpha(s,a)\}$ satisfying*

$$\sum_{s' \in S} \sum_{a \in A'_s} \alpha(s',a)[\delta_{s'}(s) - Q_{s'}(s,a)] = \ , \forall s, \quad \alpha(s,a) \geq 0, \forall s, a. \quad (10)$$

where $\delta_{s'}(s)$ is the Dirac distribution in state s'.
(ii) We have: $\mathcal{L}_\eta(U) = \mathcal{L}_\eta(U_S) = co\mathcal{L}_\eta(U_D)$ where $co\mathcal{L}_\eta(U_D)$ is the convex hull of $\mathcal{L}_\eta(U_D)$.
(iii) For any $\theta \in \mathcal{L}_\eta(U)$, define the individual stationary policy $u \in \mathcal{U}_S$ by

$$u_s(a) = \frac{\theta(s,a)}{\sum_{a \in A_s} \theta(s,a)}.$$

Then $f_{\eta,u} = \theta$.

Transforming the MDEG into a standard EG

Consider the following standard evolutionary game **EG**:

- the finite set of (pure) actions of a player is U_D,
- the fitness of a player that uses $v \in U_D$ when the other use a policy $u \in U_S$ is given by (5).
- Enumerate the strategies in U_D such that $U_D = (u_1, ..., u_m)$.
- Define $\gamma = (\gamma_1, ..., \gamma_m)$ to be a probability measure over the set U_D; where γ_i is the fraction of the population that uses u_i. γ can be interpreted as a mixed strategy which we denote by $\widehat{\gamma}$.

REMARK 4.1. *Here the convex combination $\epsilon\widehat{\gamma} + (1-\epsilon)\widehat{\gamma}'$ of the two mixed strategies $\widehat{\gamma}$ and $\widehat{\gamma}'$ is simply the mixed strategy whose ith component is given by $\epsilon\gamma_i + (1-\epsilon)\gamma'_i$, $i = 1, ..., m$.*

PROPOSITION 4.1. *(i) $\widehat{\gamma}$ is an equilibrium for the game* **EG** *if and only if it is a WESS for the original MDEG.*
(ii) $\widehat{\gamma}$ is an ESS for the game **EG** *if and only if it is a WESS for the original MDEG.*

Proof. The statements hold if we allowed for only mixed policies; indeed, they follow from Lemma 4.1 and eq. (5). We have to check that if a mixed policy is an equilibrium or a WESS when restricting to U_M then it is also an equilibrium among all policies. This in turn follows from from Lemma 4.1 and eq. (7). ■

5. CONCLUSIONS

In this paper we have studied a new class of evolutionary games which we call MDEG, where the decision of each player determine transition probabilities between individual state. We have illustrated this class of game through an energy control problem in wireless networks. We had introduced already in [6] a definition of ESS strategies in stationary policies in a particular simple MDEG in which only in one state there are decisions to be taken. If we apply directly that definition to general policies (we call this here a Strong ESS) it turns out that when abandoning the restriction to stationary policies, even in this simple model there are no ESS (except for some restricted choice of parameters that results in some pure ESS). We solved this problem by defining a weaker notion of ESS using occupation measures. We have then proposed methods to determine weak ESS.

6. ADDITIONAL AUTHORS

7. REFERENCES

[1] E. Altman, K. Avrachenkov, N. Bonneau, M. Debbah, R. El-Azouzi, and D. Sadoc Menasche. Constrained stochastic games in wireless networks. In *IEEE Globecom General Symposium*, Washington D.C., 2007.

[2] E. Altman, K. Avrachenkov, N. Bonneau, M. Debbah, R. El-Azouzi, and D. Sadoc Menasche. Constrained cost-coupled stochastic games with independent state processes. *Operations Research Letters*, 36:160–164, 2008.

[3] J.W. Weibull, *Evolutionary Game Theory*, Cambridge, MA: MIT Press, 1995.

[4] Josef Hofbauer and Karl Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press 1998.

[5] L. Shapley, *Stochastic Games*, Proceedings of the National Academy of Sciences of the USA, 39, 1953.

[6] E. Altman, Y. Hayel, *A Stochastic Evolutionary Game Approach to Energy Management in a Distributed Aloha Network*, proceedings of IEEE INFOCOM, 2008.

[7] A. I. Houston and J. M. McNamara, "Fighting for food: a dynamic version of the Hawk-Dove game", *Evolutionary Ecology*, Vol 2, pp. 51-64, 1988.

[8] A. I. Houston and J. M. McNamara, "Evolutionary stable strategies in the repeated hawk-dove game", *Behavioral Ecology*, Vol 2 No 3, pp. 219-227, 1991.

[9] J. McNamara, S. Merad, E. Collins, *The Hawk-Dove Game as an Average-Cost Problem*, Advances in Applied Probability, vol. 23, no. 4, 1991.

[10] E. Altman, *Constrained Markov Decision Process*, Chapman and Hall/CRC, 1999.

[11] Valentina Corradi and Rajiv Sarin, "Continuous Approximiations of Stochastic Evolutionary Game Dynamics", Journal of Economic Theory Volume 94, Issue 2, October 2000, Pages 163-191

[12] Dean Foster and Peyton Yong, "Stochastic Evolutionary game dynamics", *Theoreticial Population Biology*, Vol 38 No 2, October 1990.

[13] Lorens A Imhof, "Long-run behavior of the stochastic replicator dynamics", *Annals of Appl Probability*, Vol 15 No 1B, 1019-1045, 2005.

[14] L. C. M. Kallenberg (1983), *Linear Programming and Finite Markovian Control Problems*, Mathematical Centre Tracts 148, Amsterdam.

[15] Michel Benaïm, Joseph Hofbauer and William H. Sandholm, "Robust Permanence and Impermanence for the Stochastic Replicator Dynamic", available in http://members.unine.ch/michel.benaim/perso/bhsa.pdf

[16] D. J. White, *Markov Decision Processes*, J. Wiley, 1993.