Markov Decision Evolutionary Games with Expected Average Fitness

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

Y. Hayel, H. Tembine, and R. El-Azouzi
LIA/CERI, University of Avignon
339, chemin des Meinajaries Agroparc
BP 1228, F-84911 Avignon Cedex, France

Abstract

Aim: To model and characterize evolutionary games where individuals have states that are described by controlled Markov chains. The action of an individual in a local interaction with another randomly selected individual determines not only the instantaneous fitness but also its probability to move to another state. The goal of a player is to maximize its time average fitness.

Mathematical methods: The main mathematical tool is occupation measures (expected frequencies of states and actions). This tool is a central one in the theory of Markov Decision Processes. We make use of the geometric properties of the set of achievable occupation measures.

Key assumption: Under any pure stationary policy of an individual, its Markov chain has a single ergodic class of states.

Results: We define and characterize a new concept of Evolutionarily Stable Strategies based on the concept of Occupation Measures. We relate this set to the concept of Evolutionarily stable set (ESSet). We present a way to transform the new type of evolutionary games into standard ones. Applying this novel framework to energy control in wireless networks, we show existence of an Occupation Measure ESS (OMESS).

Keywords: evolutionary games, occupation measure evolutionarily stable strategy, Markov decision process, energy control in wireless networks.

Introduction

Evolutionary games have been introduced 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 (see *Fisher 1930*, *Hamilton 1963,1964*, *Maynard Smith 1972*). Central in evolutionary games is the concept of Evolutionarily Stable Strategy (ESS) introduced by *Maynard Smith & Price* in 1973. ESS 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 (*Nash, 1951*). 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. Maynard 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, that we may view as random deviations. But the assumption of a very large population tends to hide this source of randomness since that randomness tends to average out. The deterministic evolutionary games may provide an interpretation in which the deterministic game is a limit of games with finitely many players who may take random actions. Such an interpretation can be found in (*Corradi and Sarin, 2000*).

Yet, other sources 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 (*Benaïm et al, 2007; Foster & Yong, 1990; Imhof, 2005*) and references therein. The ESS can then be replaced by other notions such as the Stochastic Stable Equilibrium (*Foster & Young, 1990*) .

In this paper we introduce another class of stochastic evolutionary 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 moves 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 a MDP (Markov Decision Process) 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 (*Altman et al, 2007; 2008*) which suggests the sufficiency of stationary strategies.

A simple application of an MDEG to mobile communications has been introduced in (*Altman & Hayel, 2008a*) for the case in which individual mobile terminals 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 and adapt it to the average expected fitness criterion. We shall use the long-term average reward criterion in order optimize the expected fitness that an individual can have during its lifetime (which can be very large). The analysis of MDEG with the total expected fitness criterion has been proposed in Altman & Hayel 2008b. 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 in (*Houston and McNamara 1988,1991; McNamara et al, 1991*).

The paper is organized as follows. The next section recalls notions of equilibria in evolutionary games (EG) and the link with Nash equilibrium. We then present the model and we define the new notion of occupation measure based ESS (OMESS). We propose a method to compute the OMESS based on a transformation of the MDEG into a standard EG. Finally we apply the last method to an energy control problem and we conclude the paper.

Reminder on (standard) Evolutionary Games

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.

The basic equilibrium concept is the equilibrium strategy or a Nash equilibrium.

Definition 1 *The strategy q is a Nash equilibrium if for any strategy p,*

$$J(q, q) \geq J(p, q). \tag{1}$$

An ESSet is a set of Nash equilibria which have the following special properties (see *Cressman, 2003*).

Definition 2 *A set E of symmetric Nash equilibrium is an evolutionarily stable set (ESSet) if, for all $q \in E$, we have $J(q, p) > J(p, p)$ for all $p \notin E$ and such that $J(p, q) = J(q, q)$.*

Note that for all strategies p and p' in an ESSet E, we have $J(p', p) = J(p, p)$. The concept of ESSet is stronger than Nash equilibrium and there are some simple matrix games in which such an equilibrium set does not exist (see *Weibull, 1995, page 48 example 2.7*).

In (*Thomas, 1985*), the author defines Evolutionary Stable Sets and presents an example of ESSet containing a continuum of (Nash) equilibrium strategies, none of which can be an Evolutionarily Stable Strategy (ESS). Another example is of an ESS, a special case of an ESSet restricted to one point.

The ESSet is robust against perturbation by a strategy which is outside the ESSet, but any strategy in the set need not to be robust against perturbation by another strategy within the ESSet. ESSet is asymptotically stable for the replicator dynamic (*Cressman, 2003*). Every ESSet is a disjoint union of Nash equilibria.

The stronger notion of equilibrium from evolutionary game theory is the Evolutionarily Stable Strategy (ESS). The concept of ESSets generalize the ESS as it is a one-element ESSet. 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

$$J(q, \epsilon p + (1-\epsilon)q) > J(p, \epsilon p + (1-\epsilon)q) \tag{2}$$

Definition 3 *A strategy q is said to be ESS if for every $p \neq q$ there exists some $\bar{\epsilon}_p > 0$ such that (2) holds for all $\epsilon \in (0, \bar{\epsilon}_p)$.*

In fact, we expect that if

$$\forall p \neq q, \quad J(q,q) > J(p,q) \tag{3}$$

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

$$\forall p \neq q, \quad J(q,q) = J(p,q) \text{ and } J(q,p) > J(p,p) \tag{4}$$

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 1 *(*Hofbauer and Sigmund,1998, Theorem 6.4.1, page 63*) A strategy q is ESS if and only if $\forall p \neq q$ one of the following conditions holds:*

$$J(q,q) > J(p,q), \tag{5}$$

or

$$J(q,q) = J(p,q) \textit{ and } J(q,p) > J(p,p). \tag{6}$$

Corollary 1 *Relation (3) is a sufficient condition for q to be an ESS. A necessary condition for it to be an ESS is relation (1).*

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 $1-p$, respectively.

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.

Model for Individual player

A player arrives at some random time t_0. It has a clock that dictates the times at which interactions with other players occur. It is involved in interactions that occur according to a Poisson process with rate λ. After a random number of time periods, the player leaves the system and is replaced by another one. This will be made precise below. During the player's life time, each time the timer clocks, the player interacts with another randomly selected player.

We associate with each player a Markov Decision Process (MDP) embedded at the instants of the clocks.

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 policies is U. A general policy u is a sequence $u = (u_1, u_2, \dots)$ 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 or individual time; a player is not assumed to use policies that make use of some global clocks. A policy is an individual decision which defines the sequences of action which will be taken by the individual at each individual's clock.

- The subset of mixed (respectively pure or deterministic) policies is U_M (respectively U_D). We define also the set of stationary policies U_S where such policy does not depend on time.

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 introduce next the occupation measure corresponding to a policy u.

Let $\mathbb{P}_{\eta,u}(X_t = s, A_t = a)$ be 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. The expected fraction of time units till time t, during which a user is at state s and chooses action a, is given by:

$$f^t_{\eta,u}(s,a) = \frac{1}{t} \sum_{r=1}^{t} \mathbb{P}_{\eta,u} X_t = s, A_t = a).$$

Denote $f^t_\eta(u) := \{f^t_{\eta,u}(s,a)\}$. Define by Φ^u_η 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 $f_{\eta,u}$.

Define the expected lifetime of a player corresponding to a given η and u as $T_{\eta,u} = \sum_s f_{\eta,u}(s)$. We shall assume throughout that for a given η, $\sup_{u \in U} T_{\eta,u}$ is finite. We know from Kallenberg (1983) that $\sup_{u \in U} T_{\eta,u} = \max_{u \in U_D} T_{\eta,u}$, so the assumption is equivalent to requesting that $T_{\eta,u}$ is finite for all $u \in U_D$.

We shall assume throughout that under any pure stationary policy, S_t is unichain: it is a Markov chain that has a single ergodic class of states.

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. We denote also by $r(u; s, a)$ the immediate reward that a player receives when it is at state s and it uses action a while interacting with a player using the policy u. Then, we have

$$r(u; s, a) = \sum_{s',a'} \alpha(u; s', a') r(s, a, s', a').$$

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 conditionned on its state being S_t and its action being A_t is given by

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

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

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

where $E_{\eta,v}[R_m(u)]$ is the expectation of the reward at time m considering the initial distribution η, the policy v of an individual against the policy u of the population. 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).

Unless stated differently, we shall make throughout the following assumption.

Introduce the following assumptions.

A1: When the whole population uses a policy $u \in U_S \bigcup U_M$, 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 $t \to \infty$ and π is the stationary distribution of the chain.

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^u_η is a singleton by U^*. For $u \in U^*$, the following holds:

$$F(v,u) = \inf_{z\in\Phi^v_\eta} \sum_{s,a} z(s,a) \sum_{s',a'} f_{\eta,u}(s',a') r(s,a,s',a'), \qquad (7)$$

where Φ^v_η is the set of occupation measures corresponding to the policy v and initial distribution η. The set of occupation measures will be shown to be a polytope whose extreme points correspond to strategies in U_D (*Altman 1999*). This will allow us to transform the MDEG to a standard EG.

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

$$\Phi^v_\eta \subset \Phi^w_\eta \quad \text{implies } F(v,u) \geq F(w,u). \qquad (8)$$

This, together with the fact that for any policy u and $z \in \Phi^u_\eta$ there exists a stationary policy $v \in U^*$ satisfying $f_{\eta,v} = z$, will allow us to limit ourselves to policies in U^*.

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

$$F(v,u) = \sum_{t=1}^{\infty} E_{\eta,v}[R_t(u)] = \sum_{s,a} f_{\eta,v}(s,a) \sum_{s',a'} f_{\eta,u}(s',a') r(s,a,s',a'). \qquad (9)$$

Assumption A1 would not hold if the policy of a player could depend on the absolute time or on the behavior (i.e. the actions) of other players. For example, in the standard replicator dynamics, the policy of a player adapts to the instantaneous fitness which depends also on the actions of the other players in the population. Thus A1 does not hold there. On the other hand, since players of a given class are undistinguishable, and since the lifetime distribution of a player depends only on his local time, we may expect Assumption A1 to hold. Checking A1 is beyond the scope of the paper.

Definition 4 *We shall say that two strategies u and u' are equivalent if the corresponding occupation measures are equal. We shall write $u =_e u'$. The set of occupation measures equivalent to u is denoted by $e(u) := \{v | v =_e u\}$.*

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

Defining the Occupation Measure ESS

With the expression (9) for the fitness, we observe that we are again in the framework of standard evolutionary game model and can use Definition of Theorem 1 for the Occupation Measure based ESS (OMESS) in the MDEG:

Definition 5 *(i) A strategy $u \in U^*$ is an equilibrium for the MDEG if and only if it satisfies*

$$F(u,u) \geq F(v,u). \tag{10}$$

(ii) A strategy $u \in U^*$ is a Occupation-measure ESS (OMESS) 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: $F(u,v) > F(v,v)$.*

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

Theorem 2 *A strategy u is said to be OMESS if for every $v \neq_e u$ there exists some $\overline{\epsilon}_v > 0$ such that the following holds for all $\epsilon \in (0, \overline{\epsilon}_v)$:*

$$F(u, \epsilon u + (1-\epsilon)v) > F(u, \epsilon u + (1-\epsilon)v) \tag{11}$$

In equation (11) we use a convex combination of two policies. We delay the definition of this to the next section (see Remark 1).

The following result links between the OMESS and the ESSets.

Proposition 1 *If u is an OMESS, then $e(u)$ is an ESSet.*

Proof Let u be an OMESS. We take a measure $v \notin e(u)$. By definition of equivalent classes, one of the following condition holds:

- $F(u,u) > F(v,u)$,

- $F(u,u) = F(v,u)$ and $F(u,v) > F(v,v)$.

Thus each $w \in e(u)$ is a Nash equilibrium. The second condition implies that for every $w \in e(u)$ and $v \notin e(u)$ such that $F(w,w) = F(v,w)$, we have $F(w,v) > F(v,v)$. This implies by Definition 2 that $e(u)$ is an ESSet. ■

In the following, we show that an ESSet is a weaker notion than OMESS: a problem with no OMESS may still have a non empty ESSet.

Consider a single state s and two actions h or l. Assume that the reward does not depend on the action. two pure stationary policies are u and v where u consists on playing always h and the policy v is to play always l. Then,

- the ESSet of the Markov game is all the feasible policies,
- u and v are not in the same equivalence class,
- $F(u,w) = F(w,w),\ \forall w \in e(v)^c \neq \emptyset$. v is not an OMESS.
- the game has no OMESS

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 (*Altman and Hayel, 2008a*) 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 $1 - 1/\tau$.

The transition probabilities between energy levels which are motivated by the application of energy consumption satisfy:

- 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.

Computing the OMESS

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

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

It will turn out that the expected fitness of an individual (defined in next subsection) will depend on the strategy u of that individual only through Φ_η^u. We are therefore interested in the following characteristic of $\mathcal{L}_\eta(U)$ (see *Kallenberg, 1983; Altman, 1999*):

Lemma 1 *$\mathcal{L}_\eta(U)$ is equal to the set Q_η defined as the set of $\zeta = \{\zeta(s,a)\}$ satisfying*

$$\sum_{s'\in S}\sum_{a\in A_{s'}} \zeta(s',a)[\delta_{s'}(s) - Q_{s'}(s,a)] = \eta(s), \forall s, \quad \zeta(s,a) \geq 0, \forall s,a. \tag{12}$$

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 $\zeta \in \mathcal{L}_\eta(U)$, define the individual stationary policy $u \in \mathcal{U}_S$ by

$$u_s(a) = \begin{cases} \frac{\zeta(s,a)}{\sum_{a\in A_s}\zeta(s,a)} & \text{if } \sum_{a\in A_s}\zeta(s,a) > 0 \\ \text{arbitrary number in } [0,1] & \text{if } \sum_{a\in A_s}\zeta(s,a) = 0 \end{cases}$$

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

Transforming the MDEG into a standard EG

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

- the finite set of actions of a player is U_D,
- the fitness of a player that uses $v \in U_D$ when the others use a policy $u \in U_S$ is given by (9).
- Enumerate the strategies in U_D such that $U_D = (u_1, ..., u_m)$ where

$$m = \prod_{s\in\mathcal{S}} |\mathcal{A}_s|,$$

- Define $\gamma = (\gamma_1, ..., \gamma_m)$ 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 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 $\epsilon\gamma_i + (1-\epsilon)\gamma_i'$, $i = 1, ..., m$.*

Combining Lemma 1 with eq. (7) we obtain:

Proposition 2 *(i) $\widehat{\gamma}$ is an equilibrium for the game* **EG** *if and only if it is an OMESS for the original MDEG.*
(ii) $\widehat{\gamma}$ is an ESS for the game **EG** *if and only if it is an OMESS for the original MDEG.*

Proof. The statements hold if we allowed for only mixed policies; indeed, they follow from Lemma 1 and eq. (7). We have to check that if a mixed policy is an equilibrium or a OMESS when restricting to U_M then it is also an equilibrium among all policies. This in turn follows from from Lemma 1 and eq. (9).

■

Application to Energy Control in Wireless Networks (continued)

We pursue the example of energy control applying the latest proposition in order to obtain the OMESS for this MDEG. Indeed, we will find the OMESS for the related EG game which will be written as a matrix game with dimension 4. In order to find the equilibrium of this matrix game, we have to compute the fitness $\widetilde{F}(v,u)$ for all policies v and u. We use the renewal theorem to find the expected fitness per cycle of lifetime.

$$\widetilde{F}(v,u) = p\frac{T_{\eta,v}}{T_{\eta,v}+\tau}\frac{T_{\eta,u}}{T_{\eta,u}+\tau} + (1-p)\frac{T_{\eta,v}(h)}{T_{\eta,v}+\tau}\frac{T_{\eta,u}(l)}{T_{\eta,u}+\tau} +$$

$$\tau\frac{1}{T_{\eta,u}+\tau}\frac{T_{\eta,v}}{T_{\eta,v}+\tau} - C\frac{1}{T_{\eta,v}+\tau}$$

where $T_{\eta,v}(a)$ is the expected number of times the action a is used under the policy v starting from the initial distribution η and C is the cost for a new battery.

In a first step, we have to compute the occupation measure $f_{\eta,u}$ corresponding to each policy $u \in \{u_1, u_2, u_3, u_4\}$; for that we need the probability for a user to be in each state, at time t, using action a under policy u. At initial time $t=0$, a mobile always starts with a battery full of energy, that is $\eta = (0,0,1)$. We describe the matrix game with the four following matrices:

$$\widetilde{F}_1(u_i,u_j) = \frac{T_{\eta,u_i}}{T_{\eta,u_i}+\tau}\frac{T_{\eta,u_j}}{T_{\eta,u_j}+\tau},$$

$$\widetilde{F}_2(u_i,u_j) = \frac{T_{\eta,u_i}(h)}{T_{\eta,u_i}+\tau}\frac{T_{\eta,u_j}(l)}{T_{\eta,u_j}+\tau},$$

$$\widetilde{F}_3(u_i,u_j) = \frac{1}{T_{\eta,u_j}+\tau}\frac{T_{\eta,u_i}}{T_{\eta,u_i}+\tau},$$

and

$$\widetilde{F_4} = \begin{pmatrix} \frac{1}{X_1+X_3+\tau} & \frac{1}{X_1+X_3+\tau} & \frac{1}{X_1+X_3+\tau} & \frac{1}{X_1+X_3+\tau} \\ \frac{1}{X_1+X_4+\tau} & \frac{1}{X_1+X_4+\tau} & \frac{1}{X_1+X_4+\tau} & \frac{1}{X_1+X_4+\tau} \\ \frac{1}{X_2+X_3+\tau} & \frac{1}{X_2+X_3+\tau} & \frac{1}{X_2+X_3+\tau} & \frac{1}{X_2+X_3+\tau} \\ \frac{1}{X_2+X_4+\tau} & \frac{1}{X_2+X_4+\tau} & \frac{1}{X_2+X_4+\tau} & \frac{1}{X_2+X_4+\tau} \end{pmatrix}$$

with

$$X_1 = \frac{1}{1-Q(1,l)}, \quad X_2 = \frac{1}{1-Q(1,h)},$$

$$X_3 = \frac{1}{1-Q(2,l)}, \quad X_4 = \frac{1}{1-Q(2,h)}.$$

Then we obtain the following modified fitnesses depending on the policies in the following matrix:

$$\widetilde{F} = p\widetilde{F_1} + (1-p)\widetilde{F_2} + \tau\widetilde{F_3} - C\widetilde{F_4}.$$

The OMESS of the MDEG which model energy control behaviors in wireless networks is obtained by finding the OMESS of the standard EG with the matrix of fitnesses given by $\widetilde{F}$.

Conclusions

In this paper we have studied a new class of evolutionary games which we call MDEG, where the decisions 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 (*Altman and Hayel, 2008a*) 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 called Occupation Measure ESS (OMESS). We have then proposed methods to determine OMESS and we make the link with the ESSet notion.

Introducing individual Markov decision processes into evolutionary game has many future perspectives. A natural next step would be to extend the pairwise interaction model and its related fitness to a general population game framework that could include the Wardrop equilibrium framework from road traffic, and more generally, the generating function approach of (*Vincent and Brown, 2005*).

Acknowledgments

We wish to thank Thomas Vincent for his inspiring outstanding theohretical work on evolutionary game theory, and in particular, for his efforts of making them available to the engineering community to which the authors of this paper belong. The first author wishes to thank R. Cressman for discussions that made us aware of the ESSet concept during the ISDG symposium, 2008. This work was partially supported by the INRIA Collaborative Research on POPulations, gamE theorY and Evolution (POPEYE) as well as by the European Commission within the framework of the BIONETS project IST-FET-SAC-FP6-027748, www.bionets.org.

References

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

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

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

Altman E. and Hayel. Y., 2008b. "Stochastic Evolutionary Games", Proceedings of the 13th Symposium on Dynamic Games and Applications, 30th June-3rd July. Full version submitted to *IEEE Transactions on Automatic Control*

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

Benaïm, M., Hofbauer, J., and Sandholm, W., 2007. Robust Permanence and Impermanence for the Stochastic Replicator Dynamic, available in http://members.unine.ch/michel.benaim/perso/bhsa.pdf

Corradi, V., and Sarin, R., 2000. Continuous Approximiations of Stochastic Evolutionary Game Dynamics, *Journal of Economic Theory* Volume 94, Issue 2, October 2000, Pages 163-191

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

Filar, J., and Raghavan, T., 1984. A Matrix Game Solution of the Single-Controller Stochastic Game, *Mathematics of Operations Research*, vol. 9, no. 3.

Fisher, R. A. 1930, The Genetical Theory of Natural Selection. *Oxford: Clarendon Press.*

Foster, D., and Yong, P., 1990. Stochastic Evolutionary game dynamics, *Theoretical Population Biology*, Vol 38 No 2.

Hamilton, W. D. 1963. The evolution of altruistic behavior, *American Naturalist* 97, 354-356.

Hamilton, W. D. 1964, The genetical evolution of social behviour, I and II, *Journal of Theoretical Biology* 7, 1-52.

Hofbauer, J., and Sigmund, K., 1998. *Evolutionary Games and Population Dynamics*, Cambridge University Press.

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

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

Imhof, L., 2005. Long-run behavior of the stochastic replicator dynamics, *Annals of Appl Probability*, Vol 15 No 1B, 1019-1045.

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

Nash, J., 1951. Noncooperative Games, *Annals of mathematics*, vol. 54.

Kallenberg, L., 1983. Linear Programming and Finite Markovian Control Problems, *Mathematical Centre Tracts*, 148, Amsterdam.

Maynard Smith, J., 1972. *Game Theory and the Evolution of Fighting*, In John Maynard Smith, On Evolution (Edinburgh: Edinburgh University Press), pp.8-28.

Thomas, B., 1985. On Evolutionary Stable Sets, *Journal of Mathematical Biology*, vol. 22, pp. 105-115, 1985.

Vincent, T., and Brown, J., 2005. *Evolutionary Game Theory, Natural Selection and Darwinian Dynamics*, Cambridge University Press.

Weibull, J., 1995. *Evolutionary Game Theory*, Cambridge, MA: MIT Press.