

# Markov Decision Evolutionary Games

Eitan Altman, **Yezekael Hayel**,  
Hamidou Tembine, Rachid El Azouzi  
INRIA Sophia-Antipolis, France  
University of Avignon, France

Popeye seminar, 2008

# Plan

- 1 Introduction
- 2 Markov Decision Evolutionary Games
- 3 Application to energy management in wireless networks
- 4 Numerical illustrations
- 5 Conclusions and perspectives

# Plan

- 1 Introduction
- 2 Markov Decision Evolutionary Games
- 3 Application to energy management in wireless networks
- 4 Numerical illustrations
- 5 Conclusions and perspectives

# Evolutionary Game Theory

## Evolutionary Stable Strategy (ESS)

The ESS is characterized by a property of robustness against invaders (mutations). More specifically,

- if an ESS is reached, then the proportions of each population do not change in time.
- at ESS, the populations are immune from being invaded by other small populations.
- restriction to interactions that are limited to pairwise.

This notion is stronger than Nash equilibrium in which it is only requested that a single user would not benefit by a change (mutation) of its behavior.

ESS is robust against a deviation of a **whole fraction** of the population.

# Definitions

## ESS

- $J(p, q)$  the expected immediate payoff for an individual if it uses a strategy  $p$  when meeting another individual who adopts the strategy  $q$ .
- $K$  available strategies which are called pure strategies.

## Definition

*A strategy  $q$  is said to be an ESS if for every  $p \neq q$  there exists some  $\bar{\epsilon}_q > 0$  such that for all  $\epsilon \in (0, \bar{\epsilon}_q)$ :*

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

# Important theorem

## Theorem

*A strategy  $q$  is an ESS if and only if it satisfies*

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

*or*

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

# Markov Decision Evolutionary Games (MDEG)

## MDEG

- 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 have finite life time and take during which they participate in several local interactions.
- The actions taken by a player determine not only the immediate fitness but also the transition probabilities to its next individual state.

# Plan

- 1 Introduction
- 2 Markov Decision Evolutionary Games**
- 3 Application to energy management in wireless networks
- 4 Numerical illustrations
- 5 Conclusions and perspectives

# Model for individual player

## Individual MDP

We associate with each player a Markov Decision Process (MDP) embedded at the instants of the interactions. 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  $a$ .  $\sum_{s' \in \mathcal{S}} Q_{s'}(s, a)$  is allowed to be smaller than 1.

# Model for individual player

## Policies

Define further

- The set of policies is  $\mathcal{U}$ . 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 subset of mixed (resp. pure or deterministic) policies is  $\mathcal{U}_M$  (resp.  $\mathcal{U}_D$ ). We define also the set of stationary policies  $\mathcal{U}_S$  where such policy does not depend on time.
- $\alpha(u) = \{\alpha(u; s, a)\}$  is the fraction of the population at individual state  $s$  and that use action  $a$  when all the population uses strategy  $u$ .

# Interactions and system model

## Notations

- $r(s, a, s', b)$  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  $b$ .
- The expected immediate reward of a player in state  $S_t$  and playing action  $A_t$  at time  $t$  is given by

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

- The global expected fitness when using a policy  $v$  is then

$$F_\eta(v, u) = \sum_{t=1}^{\infty} E_{\eta, v}[R_t],$$

where  $\eta$  is the initial state distribution.

# Assumptions

## Assumptions

- A1 : the expected lifetime of a player  $T_{\eta,u}$  is finite for all  $u \in U_D$ .
- A2 : When the whole population uses a policy  $u$ , 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

$$\alpha_t(u; s, a) = \frac{f_{\eta,u}(s, a)}{T_{\eta,u}}$$

for all  $s, a$  and where  $f_{\eta,u}(s, a) = \sum_{t=1}^{+\infty} p_t(\eta, u; s, a)$  is the expected number of time units during which it is at state  $s$  and it chooses action  $a$ .

# Defining the weak (resp. strong) ESS

## Equivalent class of strategies

We shall say that two strategies  $u$  and  $u'$  are equivalent if the corresponding occupation measures are equal for all state. We shall write  $u =_e u'$ .

## Definition of the WESS (resp. SESS)

A strategy  $u$  is a weak (resp. strong) ESS, denoted by WESS (resp. SESS), for the MDEG if and only if it satisfies one of the following:

$$\text{for all } v \neq_e u \text{ (resp. } v \neq u), \quad F(u, u) > F(u, v) \quad (1)$$

$$\text{for all } v \neq_e u \text{ (resp. } v \neq u), \quad F(u, u) = F(v, u) \text{ and } F(u, v) > F(v, v) \quad (2)$$

# Transforming the MDEG into a standard EG

## MDEG into an EG

The fitness function is bilinear in the occupation measures of the players. The set of occupation measures is a polytope whose extreme points correspond to strategies in  $U_D$ .

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 other use a policy  $u \in U_S$  is given by

$$\tilde{F}(v, u) = \sum_{s,a} f_{\eta,v}(s, a) \sum_{s',a'} f_{\eta,u}(s', a') r(s, a, s', a').$$

- Enumerate the strategies in  $U_D$  such that  $U_D = (u_1, \dots, u_m)$ .
- Define  $\gamma = (\gamma_1, \dots, \gamma_m)$  where  $\gamma_i$  is the fraction of the population that uses  $u_i$ .

# Transforming the MDEG into a standard EG

## Proposition

Let  $\hat{\gamma}$  be an ESS for the game **EG**. Then it is a WESS for the original MDEG.

What about stationary policies ?

## Theorem

(i) A necessary condition for a policy  $u$  to be WESS is that

$F(u, u) \geq F(v, u)$  for all stationary  $v$ .

(ii) Assume that the following set of dynamic programming equations holds: For all state  $s \in \mathcal{S}$ ,

$$F_s(u, u) = \max_a \left[ r(u; s, a) + \sum_{s'} Q_{s'}(s, a) F_{s'}(u, u) \right]. \quad (3)$$

Then  $F(u, u) \geq F(v, u)$  for any  $v$ .

(iii) If  $\eta(s) > 0$  for all  $s$ , then the converse also holds: and (3) is equivalent to  $F(u, u) \geq F(v, u)$  for all stationary  $v$ .

# Plan

- 1 Introduction
- 2 Markov Decision Evolutionary Games
- 3 Application to energy management in wireless networks**
- 4 Numerical illustrations
- 5 Conclusions and perspectives

# Model of Energy Management in a Distributed Aloha Network

## Actions and states

- A terminal  $i$  attempts transmissions during time slots.
- At each attempt, it has to take a decision on the transmission power based on his battery energy state.
- We assume that the state can take three values:  $\{F, A, E\}$  for Full, Almost empty or Empty.
- The transmission signal power of a terminal can be High ( $h$ ) or Low ( $l$ ).
- Transmission at high power is possible only when the mobile is in state  $F$ .

# Model of Energy Management in a Distributed Aloha Network

## Aloha-type game

A mobile transmits a packet with success during a slot if:

- the mobile is the only one to transmit during this slot
- the mobile transmits with high power and all others transmitting nodes use low power

# Some notations

## Notations

- $p$  is the probability for a mobile to be the only transmitter during a slot.
- $Q_i(a)$  is the probability of remaining at energy level  $i$  when using action  $a$ .
- $\alpha$  is the fraction of the population who use the action  $h$  at any given time (situation in which the system attains a stationary regime).

# Policies and fitness

## Policies

A general policy  $u$  is a sequence  $u = (u_1, u_2, \dots)$  where  $u_i$  is the probability of choosing  $h$  if at time  $i$  the state is  $F$ . We consider only *stationary policies*,  $u_i = \beta$  for all time  $i$ .

## Fitness

Let  $R_t$  denote the number of packets (zero or one) successfully transmitted at time slot  $t$  and the *fitness* of the terminal to be given by

$$\sum_{t=1}^{\infty} R_t.$$

$V_{\beta}(i, \alpha)$  is the total expected fitness (i.e. reward or valuation) of a user given that it uses policy  $\beta$ , that it is in state  $i$  and given the parameter  $\alpha$ .

# Computing fitness and sojourn times

## State E and A

- When the level of energy is in state  $E$ , the valuation is equal to  $V(E) = 0$ .
- When the state is  $A$ , the valuation is  $V(A) = \frac{p}{1-Q_A}$  and expected time during which a mobile spends in state  $A$  is  $T(A) = \frac{1}{1-Q_A}$ .

## State F

Define the dynamic programming operator  $Y(v, a, \alpha)$  to be the *total expected fitness of an individual starting at state F*, if

- It takes action  $a$  at time 1,
- If at time 2 the state is  $F$  then the total sum of expected fitness from time 2 onwards is  $v$ .
- At each time the mobile attempts transmission, the probability that another interfering mobile uses action  $h$  is  $\alpha$ .

# Computing fitness and sojourn times

## State F

$$Y(v, l) = p + Q_F(l)v + p \frac{1 - Q_F(l)}{1 - Q_A}$$

and

$$\begin{aligned} Y(v, h, \alpha) &= \alpha(p + Q_F(h)v + (1 - Q_F(h))V(A)) \\ &\quad + (1 - \alpha)(1 + Q_F(h)v + (1 - Q_F(h))V(A)), \\ &= \alpha p + (1 - \alpha) + Q_F(h)v + p \frac{1 - Q_F(h)}{1 - Q_A}. \end{aligned}$$

# Computing fitness and sojourn times

## State F

The expected time it spends at state  $F$  is

$$T(F) = \frac{1}{1 - \beta Q_F(h) - (1 - \beta) Q_F(l)}$$

The fraction of time that the mobile uses action  $h$  is then

$$\hat{\alpha}(\beta) = \beta \frac{T(F)}{T(F) + T(A)} = \beta \frac{1 - Q_A}{2 - Q_A - \beta Q_F(h) - (1 - \beta) Q_F(l)}$$

# Computing fitness and sojourn times

## State F

The total expected utility  $V_\beta(F, \alpha)$  the mobile gains starting from state  $F$  is the unique solution of  $v = (1 - \beta)Y(v, l) + \beta Y(v, h, \alpha)$ . This gives

$$V_\beta(F, \alpha) = V(A) + \frac{p + \beta(1 - p)(1 - \alpha)}{1 - Q_F(l) + \beta(Q_F(l) - Q_F(h))}.$$

## some remarks

- aggressive policy  $\beta = 1$ ,  $V_1(F, \alpha) = V(A) + \frac{1 - \alpha(1 - p)}{1 - Q_F(h)}$ ,
- passive policy  $\beta = 0$ ,  $V_0(F, \alpha) = V(A) + \frac{p}{1 - Q_F(l)}$ ,
- $V_\beta(F, \alpha)$  is either constant or strictly monotone in  $\beta$  over the whole interval  $[0, 1]$ .

# Characterization of the ESS

## Relation with EG

- The fitness that is maximized is not the outcome of a single interaction but of the sum of fitnesses obtained during all the opportunities in the mobile's lifetime.
- The ESS can be defined using the following fitness:

$$V_{\beta}(F, \hat{\alpha}(\beta')) = J(\beta, \beta').$$

- A necessary condition for  $\beta^*$  to be an ESS is

$$\text{for all } \beta' \neq \beta^*, \quad V_{\beta^*}(F, \hat{\alpha}(\beta^*)) \geq V_{\beta'}(F, \hat{\alpha}(\beta^*)).$$

# Pure equilibrium with high power

## Theorem

*Define*

$$\Delta_h := \frac{1 - Q_F(h)}{2 - Q_A - Q_F(h)}(1 - p) - \frac{Q_F(l) - Q_F(h)}{1 - Q_F(l)}p$$

*Let  $u$  be the pure aggressive strategy that uses always  $h$  at state  $F$ .*

- (i)  $\Delta_h > 0$  is a sufficient condition for  $u$  to be an ESS.*
- (ii)  $\Delta_h \geq 0$  is a necessary condition for  $u$  to be an ESS.*

## Remarks:

- If  $Q_F(l) = Q_F(h)$ , the strategy high power is obviously an ESS.
- If  $p = 0$ , there is no benefit from transmission with high power.
- $\Delta_h > 0$  is a sufficient and necessary condition for  $u$  to be a strongly immune ESS.

# Pure equilibrium with low power

## Theorem

*Define*

$$\Delta_I := p(1 - Q_F(h)) - (1 - Q_F(l))$$

*Let  $v$  be the pure strategy that uses always  $l$  at state  $F$ .*

*(i)  $\Delta_I > 0$  is a sufficient condition for  $v$  to be an ESS.*

*(ii)  $\Delta_I \geq 0$  is a necessary condition for  $v$  to be an ESS.*

## Remarks:

- If  $Q_F(l) = Q_F(h)$ , the strategy low power is not an ESS.
- The condition for the policy  $v$  to be ESS does not depend on  $Q_A$ .
- $\Delta_I > 0$  is a sufficient and necessary condition for  $v$  to be a strongly immune ESS.

# Mixed equilibrium

## Theorem

(a) Each one of the following conditions is necessary for there to exist a Weakly Immune ESS:

- Condition (i):  $\Delta_l \leq 0$ ,
- Condition (ii):  $\Delta_h \leq 0$ ,

(b) Assume that Condition (i) and (ii) hold. Then there exists a unique weakly immune ESS given by

$$\beta^* = \frac{(\overline{Q_A} + \overline{Q_F(l)})[\overline{Q_F(l)} - p\overline{Q_F(h)}]}{\overline{Q_A}p\overline{Q_F(l)} - (\overline{Q_F(l)} - \overline{Q_F(h)})(\overline{Q_F(l)} - p\overline{Q_F(h)})}$$

## Theorem

For all  $Q_A$ ,  $Q_F(l)$ ,  $Q_F(h)$  and  $p$ , the ESS  $\beta^*$  of the stochastic evolutionary game exists and is unique.

# Price of Anarchy (equilibrium aggressiveness comparison)

## PoA

- The strategy  $\tilde{\beta}$  is globally optimal if it maximizes  $V_{\beta}(F, \hat{\alpha}(\beta))$ .
- The global optimal solution is solution of a second order polynomial function.
- We compare the optimal global solution to the ESS.

## Theorem

*ESS strategy is more aggressive than the social optimum strategy, i.e.*

$$\beta^* \geq \tilde{\beta}.$$

# Matrix Game of the EG

## Matrix Game with deterministic policies

We restrict to the deterministic policies  $u_1 = (l, l)$  and  $u_2 = (l, h)$  (use always high power in state  $F$ ). The WESS of the MDEG is the ESS of a standard EG defined by through the related matrix game:

$$\tilde{G} = \begin{pmatrix} p(X_1 + X_3)^2 & p(X_1 + X_3)(X_1 + X_4) \\ (X_1 + X_3)(p(X_1 + X_4) + (1 - p)X_4) & p(X_1 + X_4)^2 + (1 - p)X_1X_4 \end{pmatrix}$$

with

$$x_1 = \frac{1}{1 - \alpha(1, l)}, \quad x_2 = \frac{1}{1 - \alpha(1, h)}, \quad x_3 = \frac{1}{1 - \alpha(2, l)}, \quad x_4 = \frac{1}{1 - \alpha(2, h)}.$$

# ESS of the EG

## Proposition

*The ESS  $\hat{\gamma}$  exists and is unique.*

## Proposition

*Policies  $\beta^*$  and  $\hat{\gamma}$  are in the same equivalent class, i.e.*

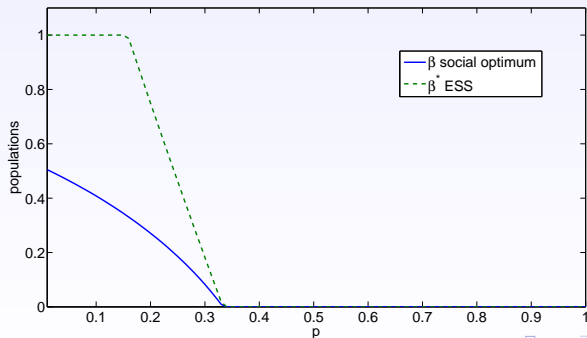
$$\beta^* =_e \hat{\gamma}.$$

# Plan

- 1 Introduction
- 2 Markov Decision Evolutionary Games
- 3 Application to energy management in wireless networks
- 4 Numerical illustrations**
- 5 Conclusions and perspectives

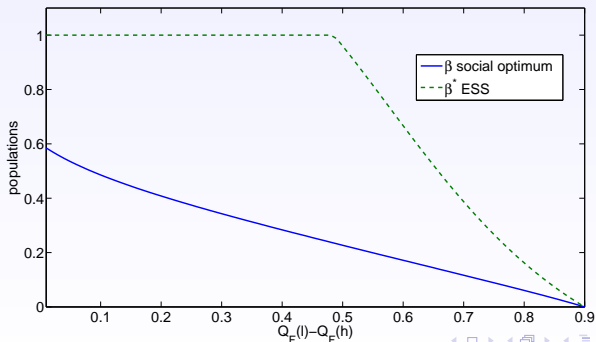
# ESS and the global optimum

Comparison of the ESS and the global optimum depending on the probability  $p$ .



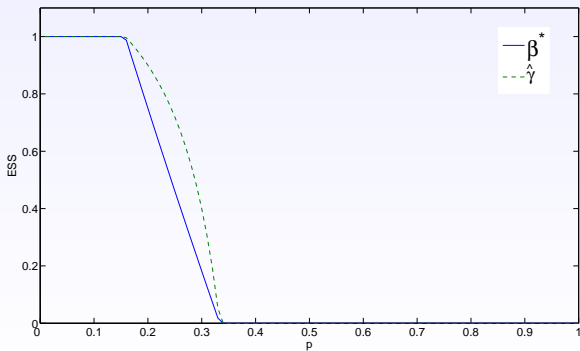
# ESS and the global optimum

Comparison of the ESS and the global optimum depending on the difference  $Q_F(l) - Q_F(h)$ .



# Comparison of the two approaches

Comparison of the two mixed ESS  $\beta^*$  and  $\hat{\gamma}$  given by the two approaches.



# Plan

- 1 Introduction
- 2 Markov Decision Evolutionary Games
- 3 Application to energy management in wireless networks
- 4 Numerical illustrations
- 5 Conclusions and perspectives**

# Conclusions

## Conclusions

- Extension of evolutionary game paradigm considering action state dependence.
- Application to competitive energy management in wireless terminals.
- Two different methods for computing ESS of a MDEG.

# Perspectives

## Perspectives

- Generalization of our energy management application.
- Develop the theoretical results to other rewards like the mean and the discounted ones.
- Notion of population dynamics into this framework.