2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Risk-sensitive mean field stochastic games

Hamidou Tembine

Abstract— **Recently, there has been much interest in understanding the behavior of large-scale systems in dynamic environment. The complexity of the analysis of large-scale systems is dramatically reduced by exploiting the mean field approach leading to macroscopic dynamical systems. Under regularity assumptions and specific time-scaling techniques the evolution of the mean field limit can be expressed in deterministic or stochastic equation or inclusion (difference or differential). In this paper, we study a risk-sensitive mean field stochastic game with discounted and total payoff criterion. We provide a risk-sensitive mean field system for the long-term total payoff and derive backward-forward mean field equations. In contrast to risk-neutral discounted case, we show the non-existence of stationary mean field response in a simple scenario with two actions for each generic player.**

I. Introduction

Mean field interactions with large number of players with different types, locations and controls can be described as a sequence of dynamic games. Since the population profile involves many players for each type, class or location, a common approach is to replace individual players and to use continuous variables to represent the aggregate average of type-location-secondary actions. The validity of this method has been proven only under specific time-scaling techniques and regularity assumptions. The *mean field limit* is then modeled by state and location-dependent time process. This type of aggregate models have been proposed in von Neumann (1944) and Nash (1951) in the mass-action interpretation. It is also known as non-atomic or population games and have been studied by Wardrop (1952, [8]) in a deterministic and stationary setting of identical players. In the infinite population games, an equilibrium m is characterized by a fixed inclusion: the support of the population profile is included in the argmax of the payoff function r,

$$\overline{\{x,\ m_x > 0\}} = \text{support}(m) \subseteq \arg\max r(m) \qquad (1)$$

In other words, if the fraction of players under a specific action is non-zero then the payoff of the corresponding action is the maximum. This large-scale methodology has inherent connections with evolutionary game theory when one is studying a large number of interacting players in different subpopulations. Different solution concepts such as evolutionarily state states or strategies, neutrally stable strategies, invadable states have been proposed and several applications can be found in evolutionary biology, ecology, control design, networking and economics (see [6], [5], [3] and the references therein).

H. Tembine is with Ecole Superieure d'Electricite, supelec, 3 rue Joliot-Curie 91192 Gif-sur-Yvette cedex, France. tembine@ieee.org

In [7], [4], risk-neutral models of interacting players in discrete time with finite number of states have been considered. The players share local resources which have finite number of states. The players are observable only through their own state which changes according to a Markov decision process. In the limit, when the number of players goes to infinity, it is found that the system can be approximated by a non-linear dynamical system.

Most formulations of discrete time mean-field models have been of risk-neutral type where the cost (or payoff, utility) functions to be minimized (or to be maximized) are the expected values of the stage-additive loss functions.

Not all behavior, however, can be captured by risk-neutral payoff functions. One way of capturing risk-seeking or risk-averse behavior is by exponentiating instantaneous payoff functions before expectation. In this paper we study discrete time risk-sensitive mean field Markov games. Our work extends the results in [2], [1] developed for risk-sensitive controlled Markov chain. Compared to the traditional Markov decision process techniques, additional difficulties arise in mean field games due to the fact the transitions probabilities may be controlled by the generic player but also by the mean field.

Our contribution can be summarized as follows. (i) we establish backward-forward equation with multiplicative mean field system, (ii) we show that if a stationary strategy is obtained maximizing the right-hand side of the mean field equation, then this strategy is best-response to mean field whenever it induces a Markov decision process with a unique positive recurrent class, however, (iii) if this last property fails, the existence of a best-response strategy cannot be generally ensured. (iv) In contrast to the risk-neutral discounted payoff where it is well-known (Shapley, 1953) that a stationary equilibrium exists under complete information and finite state and actions spaces. The result has been extended to more general state and action spaces. Here, we show that optimal stationary strategies may not exist in the risk-sensitive case. We give examples of non-existence of mean field best response and sub-optimality of stationary strategies under the risk-sensitive criterion.

The remainder of the paper is structured as follows. In the next section we overview the mean field model description. After we focus on the risk-sensitive cumulative payoff and the mean field backward forward in Section III. In section IV we analyze the risk-sensitive discounted payoff. Finally, Section V concludes the paper.

978-1-61284-799-3/11/$26.00 ©2011 IEEE

II. THE SETTING

We consider a system with n players. Time $t \in \mathbb{Z}_+$ is discrete, $\mathbb{Z}_+$ denotes the set of natural numbers. For every player j, $\mathcal{X}$ is its own-state space. In this paper $\mathcal{X}$ is finite or $\mathcal{X} = \mathbb{Z}_+$. Each individual state has two components: the type of the player and the internal state. The type is a constant during the game. The individual state of player j at time t is denoted by $x^n_{j,t} = (\theta_j, y^n_{j,t})$ where θ_j is the type and $y^n_{j,t}$ denotes the internal individual state of the player j at time t. The set of possible states $\mathcal{X}_j = \{1, 2, ..., \Theta\} \times \mathcal{Y}_j$. The set $\mathcal{Y}_j$ may include other parameters, such as classes, space location, current direction and so on. For every player j, $\tilde{\mathcal{A}}_j$ is the set of actions of that player. $\mathcal{A}_j : \mathcal{X}_j \longrightarrow 2^{\tilde{\mathcal{A}}_j}$ is a set-valued map (correspondence) that assigns to each state $x_j \in \mathcal{X}_j$ the set of actions $\mathcal{A}_j(x_j)$ that are available to player j. We assume that the set $\mathcal{A}_j(x_j)$ depends only on the type θ_j and value of the state x_j (not on the index of the player). In this paper we restrict our attention to finite number of actions per state or one-dimensional compact action set per states. The action of player j at time t is denoted by $a^n_{j,t}$. The *global state* of the system at time t is $x^n_t = (x^n_{1,t}, ..., x^n_{n,t})$. Denote by $a^n_t = (a^n_{1,t}, \dots, a^n_{n,t})$ the action profile at time t. The system x^n_t is Markovian once the action profile a^n_t are drawn under Markovian strategies. We denote the set of Markovian strategies by $\mathcal{U}$. The players are coupled not only via their instantaneous payoff function by $r^n_t(x^n_t, a^n_t)$ but also via the state evolution x^n_t i.e the evolution of $x^n_{j,t}$ may depend on the actions and states of the other players. When n is large, the stochastic game leads to a curse of dimensionality problem. Define $\tilde{M}^n_t$ to be the current population profile i.e

$$\tilde{M}^n_t(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{x^n_{j,t} = x\}}. \tag{2}$$

where $\mathbb{1}_{\{\}}$ denotes the indicator function.

At each time t, $\tilde{M}^n_t$ is in the set $\{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\}^{|\mathcal{X}|}$, and $\tilde{M}^n_t(x)$ is the fraction of players who belong to population of individual state x. The population profile M^n_t depends implicitly on the strategies adopted by the player. We denote $\mathcal{U}_j$ the set of admissible strategies of player j. For $u = (u_1, \dots, u_n) \in \prod_j \mathcal{U}_j$, and a subset $X_1 \subseteq \mathcal{X}$, define $M^n_t[u](X_1) := \frac{1}{n} \sum_{j=1}^{n} \delta_{\{x^n_{j,t}[u] \in X_1\}}$ where δ_z denotes the Dirac measure.

We assume that for any given Markovian strategy, the transition kernel L^n is invariant by any permutation of the index of the players within the same type. This implies in particular that the players are only distinguishable through their individual state. Moreover, this means that the process M^n_t is also Markovian when the sequence of Markovian strategies is given. This allows us to use existing frameworks for the weak convergence of the process M^n_t in Skhorohod topology.

Kernel definitions: Let $\mathcal{F}^n_t = \sigma(x^n_{t'}, a^n_{t'}, \ t' \leq t)$ be the filtration generated by the sequence of states and actions up to t. The evolution of the system depends on the decision of the interacting players. Given a history $h_t = (x^n_0, a^n_0, \dots, x^n_t, a^n_t)$.

$L^n(x'; x, u)$ is the transition kernel on $\mathcal{X}^n$ under the strategy u^n. The system evolves according to the kernel

$$\mathcal{L}^n(m'; m, u) := \mathbb{P}(M^n_{t+1} = m' \ | M^n_t = m, u^n_t = u)$$

where $\tilde{h}_t = (x^n_{t'}, a^n_{t'}, \ t' \leq t, x^n_t = x^n)$, such that $\frac{1}{n} \sum_{j=1}^{n} \delta_{x^n_j} = m$. $\mathcal{L}^n(m'; m, u)$ corresponds to the projected kernel of L^n. We examine the discrete time mean field limit case. Sufficient conditions for weak convergence of the process $\{M^n_t\}_{0 \leq t \leq T}$ can be found in [5], [4].

A. Classical mean field payoff

A history of length t for a generic player is a collection $(x^n_{j,t'}, a^n_{j,t'}, M^n_{t'}, \ t' \leq t)$. We denote by $\mathcal{H}_{j,t}$ the set of histories up to t. A strategy of player j is a collection of mappings from $\cup_{t \geq 0} \mathcal{H}_{j,t} \longrightarrow \Delta(\mathcal{A})$. A initial state $x^n_{j,0}$, the strategy profile σ, and a trajectory $\{M^n_t(.)\}_{t \geq 0}$ generates a probability measure $\mathbb{P}_{\sigma, x_0, m_0}$ over the set of play. We write $\mathbb{E}_{\sigma, x_0, m_0}$ to denote the expectation operator under $\mathbb{P}_{\sigma, x_0, m_0}$. The long-term total payoff is given by $R_\infty(\sigma, x, m) = \mathbb{E}_{\sigma, x, m} \sum_{t \geq 0} r_t(x_t, a_t, m_t)$, the $\delta-$discounted payoff is $R_\delta(\sigma, x, m) = \mathbb{E}_{\sigma, x, m} \sum_{t \geq 0} \delta^t (1 - \delta) r_t(x_t, a_t, m_t)$, and the limiting time-average payoff (also called Cesaro mean payoff) is given by $R(\sigma, x, m) = \liminf_T$

$$\mathbb{E}_{\sigma, x_0, m_0} \left(\frac{1}{T+1} \sum_{t=0}^{T} r_t(x_t, a_t, m_t) \ | x_0 = x, m_0 = m \right).$$

Recall that the mean field system in the risk-neutral cases are given by *Bellman-Shapley* equations (backward for finite horizon, fixed-point for infinite horizon) and *Poisson* equation (for the time-average). In this paper, we establish the analogous of these equations under the mean field risk-sensitive payoffs.

B. Risk-sensitive formulations

A link between stochastic and deterministic mean field viewpoints is provided by considering risk-sensitive stochastic approach. Let $g(y)$ be a smooth function such that $g'(y) > 0, \ g''(y) \neq 0$. The risk-sensitive approach consists to optimize the expectation $\mathbb{E}(g(R))$ where R is the traditional long-term payoff function. The certainty-equivalent expectation $e(R)$ is defined by $g(e(R)) = \mathbb{E}(g(R))$. When $g = e^{\frac{y}{\mu}}$ is exponential,

$$e(R) = g^{-1}\left(\mathbb{E}(g(R))\right) = \mu \log \left(\mathbb{E}\left(e^{\frac{R}{\mu}} \right) \right), \ \mu > 0.$$

The case where μ is negative can be examined following in a similar way. These equalities can be interpreted as follows. A generic player with payoff criterion g is indifferent between the random (thus uncertain) payoff R and the (certain) payoff $e(R)$. We define

$$F_{\delta,\mu}(\sigma, x, m) = \mu \log \mathbb{E}\left(e^{\frac{1}{\mu} \sum_{t \geq 0} \delta^t (1-\delta) r_t(x_t, a_t, m_t)} \right) \tag{3}$$

$$= g^{-1}(\mathbb{E}(g(R_\delta))) \tag{4}$$

$$F_{\infty,\mu}(\sigma, x, m) = \mu \log \mathbb{E}\left(e^{\frac{1}{\mu} \sum_{t \geq 0} r_t(x_t, a_t, m_t)} \right) \tag{5}$$

III. Optimality for the risk-sensitive payoff

We first observe that if we translate $r_t(.)$ with a constant c, $r_t(.)+c$ then the risk-sensitive payoff $F_{\infty,\mu}$ becomes $F_{\infty,\mu}+c$. This means that the optimal strategies and the equilibrium strategies are unchanged by the translation operation. Thus, we can consider positive function $r_t(.)$.

Assumption A0: The mapping $r_t(.)$ is positive. The infinite sum in $F_{\infty,\mu}$ is finite. In the continuous case, the payoff and the transition probabilities are continuous in (a, m).

Define $v_{j,\mu}(x,m) = \sup_\sigma F_{\infty,\mu}(\sigma, x, m)$. Let $q_{x\sigma x'}(m)$ be the marginal of the limiting of L^n for a generic player.

A. Risk-sensitive Mean field

We establish a risk-sensitive mean field equation for the function $v_{j,\mu}$.

Proposition 1: The optimal value $v_{j,\mu}(x,m)$ of a generic player j with type θ_j corresponding to the first component of x satisfies the $\mu-$risk sensitive mean field best-response equation: $g(v^*_{j,\mu}(x_t,m_t)) =$

$$\sup_{a\in\Delta(\mathcal{A}(x_t))}\left[e^{\frac{1}{\mu}r_t(x_t,a,m_t)}\sum_{x'}q_{x_tax'}(m_t)g(v^*_{j,\mu}(x',m_{t+1}))\right]$$

Proof: See appendix. ∎

Next, we define an equilibrium concept called *mean field equilibrium* for this class of dynamic games.

Definition 1: A pair $(u^*_t, m^*_t)_{t\geq 0}$ is a risk-sensitive mean field equilibrium if $\{u^*_t\}_{t\geq 0}$ is a best-response to the mean field trajectory $\{m^*_t\}_{t\geq 0}$ and for any time t, u^*_t generates the mean field m^*_t.

We are now prepared to state the risk-sensitive mean field system.

Corollary III-B: Assume $A0$ holds. Then, a risk-sensitive mean field system is obtained i.e the optimal value and the associated mean field satisfy:

$$\begin{cases} g(v^*_{j,\mu}(x_t,m_t)) = \max_{u\in\mathcal{A}_j(x_t)}\left[e^{\frac{1}{\mu}r_t(x_t,u,m_t)}\right. \\ \left.\sum_{x'}q_{x_tux'}(m_t)g(v^*_{j,\mu}(x',m_{t+1}))\right] \\ m_{t+1}(x') = \sum_{\bar{x}\in\mathcal{X}} m_t(\bar{x})\mathcal{L}(x'|\bar{x},u^*_t,m_t) \end{cases}$$

where

$$u^*_t \in \arg\max_u e^{\frac{1}{\mu}r_t(x_t,u,m_t)}\sum_{x'}q_{xux'}(m_t)g(v^*_{j,\mu}(x',m_{t+1})).$$

Proof: The proof of this result follows from Proposition 1 and the consistency relationship between the played actions by the individual players and the resulting mean field limit generating by the transition kernel. ∎

Following the same reasoning, a stationary mean field equilibrium (u^*, m^*) should satisfy

$$\begin{cases} g(v^*_{j,\mu}(x,m^*)) = \max_{u\in\Delta(\mathcal{A}_j(x))}\left[e^{\frac{1}{\mu}r(x,u,m^*)}\times\right. \\ \left.\sum_{x'}q_{xux'}(m^*)g(v^*_{j,\mu}(x',\sum_{\bar{x}\in\mathcal{X}}m^*(\bar{x})\mathcal{L}(x'|\bar{x},u^*,m^*)))\right] \\ = e^{\frac{1}{\mu}r(x,u^*(x),m^*)}\sum_{x'}q_{xu^*(x)x'}(m^*)g(v^*_{j,\mu}(x',m^*) \\ m^*(x') = \sum_{\bar{x}\in\mathcal{X}}m^*(\bar{x})\mathcal{L}(x'|\ \bar{x},u^*,m^*) \\ r_t(.) = r(.). \end{cases}$$

where $u^* \in \arg\max_u \left\{e^{\frac{1}{\mu}r(x,u,m^*)}\sum_{x'}q_{xux'}(m^*)\right.$ $\left.g(v^*_{j,\mu}(x',\sum_{\bar{x}\in\mathcal{X}}m^*(\bar{x})\mathcal{L}(x'|\bar{x},u^*_t,m^*)))\right\}$. To begin with the notion of irreducibility in stochastic games, recall that a set S is closed under σ if $\sum_{x'\in S}q_{x\sigma x'}(m) = 1$ for every x, whereas is communicating under σ if for each $x\in S$ there exists a positive integer $n(x,x',\sigma) = n$ such that $\mathbb{P}(x_t = x'|x_0 = x, m, \sigma) > 0$; a subset $S\subset\mathcal{X}$ is a recurrent class if S is both closed and communicating, so that two different recurrent classes are disjoint. We say that the Markov decision process with unique positive recurrent class under a strategy $u\in\mathcal{U}$ if the corresponding Markov chain has a unique positive recurrent class. This property is referred as *unichain property*. For finite state case, positive recurrent refers to the case where the expected return time is finite. Below we give sufficient conditions for a stationary strategy π to be best response to m^*.

Proposition 2: Assume $A0$ holds. Assume a stationary strategy π satisfies:

- $\forall x,\quad g(v^*_{j,\mu}(x,m^*)) = e^{\frac{1}{\mu}r(x,\pi(x),m^*)}\sum_{x'}q_{x\pi(x)x'}(m^*)g(v^*_{j,\mu}(x',m^*))$,
- The strategy π generates a Markov decision process with unique positive recurrent class,
- $m^*(x') = \sum_{\bar{x}\in\mathcal{X}}m^*(\bar{x})\mathcal{L}(x'|\bar{x},\pi(\bar{x}),m^*)$

Then, π is a risk-sensitive compatible best-response to the mean field (among all the general strategies).

Proof: A sketch proof is provided in Appendix. ∎

We now state a general comparison result for the optimal value for the best response to mean field.

Proposition 3: Suppose that $A0$ holds. Let $v' : \mathcal{X}\times\Delta(\mathcal{X})\longrightarrow[0,\infty)$ be a function such that for all (x_t,m_t),

$$g(v'(x_t,m_t)) \geq \sup_{a\in\mathcal{A}(x_t)} e^{\frac{1}{\mu}r_t(x_t,a,m_t)}\sum_{x'}q_{x_tax'}(m_t)g(v'_\mu(x',m_{t+1})).$$

Then, the optimal value v^*_μ is bounded by v' i.e $v'\geq v^*_\mu = \sup_\sigma F_{\infty,\mu}(\sigma,x,m)$.

Proof: In appendix. ∎

C. Non-existence if not unichain

In this subsection we examine the non-existence of optimal response if the assumptions of unique positive recurrent class fail. Suppose $\mu > 0$. Our example has two individual states $\mathcal{X} = \{0,1\}$. The set of actions in state 0 is reduced to a singleton $\mathcal{A}_j(0) = \{0\}$, and the set of actions in state 1 is $\mathcal{A}_j(1) = [0,1]$ (a continuum set of actions). The payoff in state 0 is 0 and the payoff is state 1 is given by a reward of 1 and a cost of investment proportional to a. We choose a normalized factor $\frac{\mu}{2}$. Thus, the payoff in state 1 is $r(1,a,m) = \frac{\mu}{2} - a\frac{\mu}{2}$. The transition probabilities in state 0 are $q_{000}(m) = 1, q_{001} = 0$. The transitions from state 1 are $q_{1a1}(m) = a = 1 - q_{1a0}(m)$. For $a < 1$, the more the player investment is high the more he/she has chance to stay in state 1 but has a higher cost (of investment). For $a = 1$, the payoff is zero and the state will move to 0 the next slot.

Thus, for every strategy σ, the state 0 is an absorbing state. One has, $\mathbb{P}\left(\sum_{t\geq 0}r(x_t,a_t,m_t) = 0 \mid x_0 = 0\right) = 1$.

Thus, the expectation

$$\mathbb{E}_{\sigma,x,m}\left[g\left(\sum_{t\geq 0} r(x_t,a_t,m_t)\right) \mid x_0=0\right]=0.$$

and payoff under σ is zero if the starting state is 0. Hence, $v^*_\mu(0,m)=0$, for any m. Then, the analysis reduces to the events until absorption i.e the exit time from state 1. A pure stationary strategy in this mean field stochastic game consists to specify the action to be played in state 1 (because in state 0 the only available choice is 0). We consider the stationary strategy π defined by $\pi(0)=0,\ \pi(1)=a_1$

Proposition 4:
- The payoff is monotone in a_1.
- There is an optimal payoff. The optimal payoff is $\mu\log 2$.
- There is no stationary strategy that is best response to mean field.

Since action spaces are compact in any state, there exists a stationary strategy solving the dynamic programming equation. The above result shows that such strategies are not optimal for the risk-sensitive payoff. This discontinuity comes from the discontinuity in the transition probabilities in the class of reduction of chain.

This example tell us that there is a big difference between standard repeated games (playing the game all the steps or equivalently single state stochastic game) and stochastic games (in the sense of Shapley 1953). In standard repeated games, playing an optimal strategy at each step leads to an optimal strategy for the long-term game. Here, it is not the case. It is important to notice that the strategy π provides an $\epsilon-$*optimal response to the mean field* for arbitrary small $\epsilon>0$. However, there is no $0-$optimality.

IV. Discounted payoff

Following the above analysis, we established that the $\delta-$discounted value satisfies:

$$\begin{cases} g(v^*_{j,\mu,\delta}(x_t,m_t))=\max_{u\in\mathcal{A}_j(x_t)}\left[e^{\frac{1-\delta}{\mu}r_t(x_t,u,m_t)}\right. \\ \left.\sum_{x'} q_{x_t u x'}(m_t) g(\delta v^*_{j,\mu,\delta}(x',m_{t+1}))\right] \\ m_{t+1}(x')=\sum_{\bar{x}\in\mathcal{X}} m_t(\bar{x})\mathcal{L}(x'|\bar{x},u^*_t,m_t) \end{cases}$$

It is well known that for the risk-neutral case, under irreducibility conditions of Markov decision process, that when the discounted goes to zero the limitting values (resp. the optimal strategies in the discounted case give optimal value (resp. optimal strategies) of the average cost criterion. The same property holds for a fixed mean field trajectory.

It is natural to ask whether a vanishing discount approach is possible or not in the risk-sensitive mean field case.

As we will see in the next sections, the answer to this question is negative for the risk-sensitive payoff.

A. Non-existence

In this subsection, we provide a non-existence result for stationary strategies under the payoff F_δ. There is no type and no resource state (equivalently both are singletons). Let $\mathcal{X}$ be the set of natural numbers. There is only one choice in state 0, $\mathcal{A}(0)=\{a_0\}$. There are two actions in any state $x\geq 1$, $\mathcal{A}(x)=\{a_0,a_1\},\ \forall x\geq 1$. The transition probabilities of a generic player j, $L^n(x^n_{j,t+1};x^n_{j,t},a^n_{j,t},m^n_t)$ is given by $L^n(0;0,a_0,m^n)=1=L^n(0;x,a_0,m^n),\ \forall x\geq 1,\ L^n(x;x,a_1,m^n)=\frac{1}{2}-\epsilon\xi^n(m^n),\ L^n(0;x,a_1,m^n)=\frac{1}{2}+\epsilon\xi^n(m^n),\ \forall x\geq 1$. When $n\longrightarrow\infty$, the term $\xi^n\longrightarrow 0$. The asymptotic of the payoff $r(x^n_{j,t},a^n_{j,t},m^n_t)$ is given by $r(x,a_0,m)\sim\chi_1=\frac{2}{3}\delta_0+\frac{1}{3}\delta_3$, $r(x,a_1,m)\sim\chi_{2,x}$, where $\chi_{2,x}=\bar{\chi}_{2,t_x}$, $\bar{\chi}_{2,t}\sim\frac{1}{2}\delta_0+\frac{1}{2}\delta_{2+\frac{1}{t}},\ t\geq 1$.

Define $\gamma_1=\max\{\frac{1}{\mu_k},\ \frac{1}{\mu_k}\leq\delta(1-\delta)\}$. This is well-defined because $\mu_k\longrightarrow+\infty$. $\gamma_l=\max\{\frac{1}{\mu_k},\ \frac{1}{\mu_k}\leq\delta^l(1-\delta),\ \frac{1}{\mu_k}<\gamma_{l-1}\},\ l\geq 2$. Let $t_l=\max\{t,\ \delta^t(1-\delta)\geq\gamma_l\}$. We choose $\chi_1,\ \chi_2$ independent.

$$\mathbb{E}\left[e^{\delta^t(1-\delta)r(x_t,a_0,m_t)}\right]=\mathbb{E}\left[e^{\delta^t(1-\delta)\chi_1}\right]$$

$$\mathbb{E}\left[e^{\delta^t(1-\delta)r(x_t,a_1,m_t)}\right]=\mathbb{E}\left[e^{\delta^t(1-\delta)\chi_2}\right]$$

Now, using the property of the generating function of the random variable χ_1,χ_2, one has $\mathbb{E}\left[e^{\delta^t(1-\delta)r(x,a_1,m_t)}\right]\leq\mathbb{E}\left[e^{\delta^t(1-\delta)r(x,a_0,m_t)}\right],\ t\leq t_x,\ x\geq 1$ Similarly,

$$\mathbb{E}\left[e^{\delta^t(1-\delta)r(x,a_1,m_t)}\right]>\mathbb{E}\left[e^{\delta^t(1-\delta)r(x,a_0,m^n_t)}\right],\ t>t_x,\ x\geq 1$$

Based on this observation, we construct the time-dependent strategy $\sigma=(\sigma^*_t)_{t\geq 0}$.

$$x\in\mathbb{Z}_+,\ \sigma^*_t(x)=\begin{cases} a_1 \text{ if} & t>t_x \\ a_0 \text{ if} & t\leq t_x \end{cases}$$

Proposition 5: (i) The strategy σ^*_t is an optimal strategy in response to the mean field. (ii) No stationary strategy can be optimal.

Proof: The statement (i) follows from the fact the difference $\mathbb{E}\left(e^{\frac{1}{\mu}\chi_{2,x}}-e^{\frac{1}{\mu}\chi_1}\right)$ is $\frac{1}{2}+\frac{1}{2}e^{(2+\frac{1}{x})\frac{1}{\mu}}-2/3-1/3e^{\frac{3}{\mu}}$ is zero for some $\mu=\mu^*_x$, strictly positive for $\mu>\mu^*_x$ and strictly negative for $\mu<\mu^*_x$. Moreover the mapping (integers) $x\longrightarrow\mu^*_x$ is strictly increasing and goes to infinity when x goes to infinity. The second statement follows from the fact any strategy is weakly dominated by σ^*_t. Since $t_x\longrightarrow+\infty$ when $x\longrightarrow+\infty$, σ^*_t strictly dominates any stationary strategy. ■

S-modular risk-sensitive mean stochastic games

We provide sufficient conditions for structural results, monotonicity of optimal response to mean field and the associated value functions. The monotonicity can be the state variable or time t. Since one has a multiplicative Bellman-Shapley equation, we need to have properties that preserve the S-modularity of a product of two functions. In order words, what are the conditions under which g_1g_2 is S-modular? Let $(E,\preceq)$ be a lattice. We say that g_1 is sub-additive if $g_1(\inf(e,e'))+g_1(\sup(e,e'))\leq g_1(e)+g_1(e')$. Here we report a well-known result.

Proposition 6: Let $(E,\preceq)$ be a lattice and g_1,g_2 be non-negative and sub-additive functions on E. Assume in addition that for any non-comparable pair $(e,e')\in E^2$, such that

$g_1(e) \leq g_1(e')$, one has $g_1(e) = g_1(inf(e,e'))$, $g_2(e') = g_2(sup(e',e))$. Then the product $g_1 g_2$ is sub-additive in E.

We apply this result to $g_1 = e^{\frac{1}{\mu} r(x,a,m)}$ and $g_2 = \sum_{x'} q_{xax'}(m) g(v(x', \langle m, \mathcal{L} \rangle))$. One can use Tarski's fixed-point theorem to establish the existence of fixed point under suitable conditions.

V. Conclusion

In this paper we have presented risk-sensitive mean field stochastic games as well as their optimality equations. We provided examples of non-existence and suboptimality of stationary strategies.

References

[1] A. Brau-Rojaa, R. Cavazos-Cadena, and E. Fernandez-Gaucherand. Controlled markov chains with risk-sensitive average cost criterion: some counterexamples. *Proceedings of the 37th IEEE Conference on Decision & Control, Tampa, Florida USA*, December 1998.

[2] R. Cavazos Cadena and R. Montes-De-Oca. Optimal stationary policies in risk-sensitive dynamic programs with finite state space and nonnegative rewards. *Applicationes Mathematicae*, 27:167–185, 2000.

[3] H. Tembine. Distributed strategic learning for wireless engineers. *Lecture notes, 300 pages, Supelec*, June 2010.

[4] H. Tembine. Mean field stochastic games: convergence, q/h learning, optimality. *American Control Conference, pp. 2423 - 2428, ACC, San Francisco, California, US.*, June 2011.

[5] H. Tembine. Mean field stochastic games: Simulation, dynamics and applications. *Lecture notes, 350 pages, Supelec*, July 2011.

[6] H. Tembine, E. Altman, R. ElAzouzi, and Y. Hayel. Evolutionary games in wireless networks. *IEEE Trans. on Systems, Man, and Cybernetics, Part B, Special Issue on Game Theory*, June 2010.

[7] H. Tembine, J. Y. Le Boudec, R. ElAzouzi, and E. Altman. Mean field asymptotic of markov decision evolutionary games and teams. *in the Proc. of GameNets*, May 2009.

[8] J. G. Wardrop. Some theoretical aspects of road traffic research. *Proceedings, Institute of Civil Engineers, PART II*, 1, 1952.

Appendix

For the clarify the presentation, the proofs are done for time-homogeneous instantaneous payoff function i.e $r_t(.) = r(.)$.

Proof of Proposition 1 First, observe that $g(\sum_{t\geq 0} r(x_t,a_t,m_t)) = e^{\frac{r(x_0,a_0,m_0)}{\mu}} g(\sum_{t\geq 1} r(x_t,a_t,m_t))$, for any arbitrary strategy σ, $x_0, x_1 \in \mathcal{X}$, $a_0 \in \mathcal{A}(x)$, the Markov property implies that

$$G_\infty = \mathbb{E}\left(g(\sum_{t\geq 0} r(x_t,a_t,m_t) \mid x_0,a_0,m_0,x_1,m_1)\right)$$

$$= e^{\frac{r(x_0,a_0,m_0)}{\mu}} \mathbb{E}_\sigma \left(g(\sum_{t\geq 1} r(x_t,a_t,m_t) \mid x_0,a_0,m_0,x_1,m_1)\right)$$

$$= e^{\frac{r(x_0,a_0,m_0)}{\mu}} \mathbb{E}_{\sigma'} \left(g(\sum_{t\geq 0} r(x_t,a_t,m_t) \mid x_1,m_1)\right)$$

where σ' is the strategy induced by σ after the history (x_0,a_0,m_0) i.e $\sigma'(.|h_t) = \sigma(.|x_0,a_0,m_0,h_t)$. Since g is positive and increasing,

$$e^{\frac{r(x_0,a_0,m_0)}{\mu}} g(F_{\infty,\mu}(\sigma',x_1,m_1)) \leq e^{\frac{r(x_0,a_0,m_0)}{\mu}} g(v^*_\mu(x_1,m_1)).$$

Now, we taking the expectation with respect to x_1 yields

$$\mathbb{E}_\sigma \left(g(\sum_{t\geq 0} r(x_t,a_t,m_t) \mid x_0,a_0,m_0)\right)$$

$$\leq e^{\frac{r(x_0,a_0,m_0)}{\mu}} \sum_{x_1} q_{x_0a_0x_1}(m_0) g(v^*_\mu(x_1,m_1))$$

$$\leq \sup_a \left[e^{\frac{r(x_0,a,m_0)}{\mu}} \sum_{x_1} q_{x_0ax_1}(m_0) g(v^*_\mu(x_1,m_1))\right].$$

Taking the expectation with the respect to a_0 gives

$$g(F_\mu(\sigma,x_0,m_0)) = \mathbb{E}_\sigma \left(g(\sum_{t\geq 0} r(x_t,a_t,m_t) \mid x_0,m_0)\right)$$

$$\leq \sup_{a\in\mathcal{A}(x_0)} \left[e^{\frac{r(x_0,a,m_0)}{\mu}} \sum_{x_1} q_{x_0ax_1}(m_0) g(v^*_\mu(x_1,m_1))\right]$$

Since the strategy σ is arbitrary and g is increasing and continuous, $g(v^*_\mu(x_0,m_0))$

$$\leq \sup_{a\in\mathcal{A}(x_0)} \left[e^{\frac{r(x_0,a,m_0)}{\mu}} \sum_{x_1} q_{x_0ax_1}(m_0) g(v^*_\mu(x_1,m_1))\right].$$

Now, we establish the reverse inequality. Fix $\epsilon > 0$. For all $x \in \mathcal{X}$, select an action $a_x \in \mathcal{A}(x)$ and a strategy satisfying $F^*_\mu(\sigma_x,x,m) \geq v^*_\mu(x,m) - \epsilon$ (the existence of such strategy follows from the definition of sup). Now we construct a new strategy as follows: For each state x, $\tilde{\sigma}_0(a_x|x) = 1$, $\tilde{\sigma}_t(.|h_t) = \sigma_{t-1}(.|x_t,a_{t-1},\ldots,x_2,a_1,x_1)$. Let us compute the payoff under $\tilde{\sigma}$.

$$\mathbb{E}_\sigma \left(g(\sum_{t\geq 0} r(x_t,a_t,m_t)) \mid x,a_x,m,x_1\right) =$$

$$e^{\frac{1}{\mu} r(x,a_x,m)} \mathbb{E}_{\tilde{\sigma}} \left(g(\sum_{t\geq 0} r(x'_t,a_t,m_t)) \mid x'_0 = x_1\right)$$

$$= e^{\frac{1}{\mu} r(x,a_x,m)} g(F_\mu(\tilde{\sigma},x_1)) \geq e^{\frac{1}{\mu} r(x,a_x,m)} g(v^*_\mu(x_1) - \epsilon) =$$

$$e^{\frac{-\epsilon}{\mu}} e^{\frac{1}{\mu} r(x,a_x,m)} g(v^*_\mu(x_1)).$$

Taking the expectation with the respect to x_1 and a_x, one gets $g(F_\mu(\tilde{\sigma},x)) \geq$

$$e^{\frac{-\epsilon}{\mu}} e^{\frac{1}{\mu} r(x,a_x,m)} \sum_{x_1} q_{xa_xx'}(m) g(v^*_\mu(x_1,m_{t+1})).$$

Since $g(v^*_\mu(x,m)) \geq g(F_\mu(\sigma,x,m))$ for any σ, the above inequality implies that $g(v^*_\mu(x_t,m_t)) \geq$

$$e^{\frac{-\epsilon}{\mu}} e^{\frac{1}{\mu} r(x_t,a_{x_t},m_t)} \sum_{x_1} q_{x_ta_{x_t}x'}(m_t) g(v^*_\mu(x',m_{t+1}))$$

for arbitrary $\epsilon > 0$ and $a_{x_t} \in \mathcal{A}(x_t)$. Thus, $g(v^*_\mu(x_t,m_t)) \geq$

$$\sup_a e^{\frac{1}{\mu} r(x_t,a,m_t)} \sum_{x'} q_{x_tax'}(m_t) g(v^*_\mu(x',m_{t+1}))$$

where m_{t+1} is driven by $\mathcal{L}$. Combining together one gets the announced result.

PROOF OF PROPOSITION 3

$g(v'(x,m)) \geq \mathbb{E}_{\sigma,x_1}(e^{\frac{1}{\mu}r(x_0,a_0)}g(v'(x_1,m_1)) \mid x_0 = x, m_0 = m)$. We use the Markov property and induction method to prove the comparison inequality. $g(v'(x_t,m_t)) \geq \mathbb{E}_{\sigma,x_{T+1}}(e^{\frac{1}{\mu}\sum_{t=0}^{T} r(x_t,a_t,m_t)}g(v'(x_{T+1},m_{T+1})) \mid x_t = x, m_t = m)$. Since $v' \geq 0$ and g is increasing, $g(v'(x_{t+1},m_{t+1})) \geq g(0)$

$$g(\sum_{t=0}^{T} r(x_t,a_t,m_t)) \longrightarrow g(\sum_{t=0}^{\infty} r(x_t,a_t,m_t)) < +\infty$$

by positivity of $r()$ and assumption $A0$.

Thus, the monotone convergence theorem implies that

$$g(v'(x,m)) \geq \mathbb{E}_{\sigma}\left(g(\sum_{t\geq 0} r(x_t,a_t,m_t)) \mid x_0 = x, m_0 = m\right)$$

Hence, $g(v'(x,m)) \geq g(F_{\mu}(\sigma,x,m))$, for all σ, x, m. This completes the proof.

PROOF OF PROPOSITION 2

We provide only a sketch of proof. Assume $A0$. Let π be a stationary strategy satisfying

(a) $\forall x$, $g(v^*_{j,\mu}(x,m^*)) = e^{\frac{1}{\mu}r(x,\pi(x),m^*)}\sum_{x'} q_{x\pi(x)x'}(m^*)g(v^*_{j,\mu}(x',m^*))$,

(b) The strategy π generates a Markov decision process with unique positive recurrent class,

(c) $m^*(x') = \sum_{\bar{x}\in\mathcal{X}} m^*(\bar{x})\mathcal{L}(x'|\bar{x},\pi(\bar{x}),m^*)$

For each x, m and t, let $w_t(x,m)$ be the equivalent of $v^*_{j,\mu}(x_t,m_t)$ with the respect to g starting from t (not from 0).

$$g(w_t(x_t,m_t)) = \mathbb{E}_{\pi}\left[g(v^*_{j,\mu}(x_t,m_t)) \mid x_0 = x, m_0 = m\right].$$

The key element for the optimality of π as best response to m is the ergodic Markov theorem which will give the independence of payoff in x. There exists a positive function $c(.)$ which depends only on m such that

$$\lim_{t\longrightarrow\infty} w_t(x,m) = c(m),\ \forall x.$$

Let prove this statement. Since π satisfies the relation (b) the Markov property yields for a random variable x_t

$$(*)\ g(v^*_{j,\mu}(x_t,m_t)) =$$
$$\mathbb{E}_{\pi}\left(e^{\frac{1}{\mu}r(x_t,a_t,m_t)}g(v^*_{j,\mu}(x_{t+1},m_{t+1})) \mid x_t, m_t\right).$$

and $g(w_{t+1}(x,m)) = \mathbb{E}_{\pi}(g(w_t(x_1)) \mid x_0 = x)$. Then, $(*)$ implies that

$$g(v^*_{j,\mu}(x_t,m_t)) \geq \mathbb{E}_{\pi}(g(v^*_{j,\mu}(x_{t+1},m_{t+1})) \mid x_t)$$

(because the positivity of μ and $r(.)$ implies that the term exponential in $(*)$ is greater than 1). This gives the inequality $g(w_{t+1}(x,m)) \leq g(w_t(x,m)),\ \forall\ x$. i.e $t \longrightarrow w_t$ is monotone decreasing in time. Note that $w_0(x,m) = v^*_{j,\mu}(x,m)$. There exists a function $w^*(x,m)$ such that $w_t(x,m)$ converges to $w^*(x,m)$. We now use the unichain property which implies that there is a unique invariant distribution $(i_{x'}(m))_{x'\in\mathcal{X}}$ of the Markov decision process induced by π and m such that,

$$g(w^*(x,m)) = \lim_{T\longrightarrow\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\pi}\left(g(w^*(x_t,m_t))|x,m\right)$$
$$= \sum_{x'} i_{x'}(m)g(w^*(x',m)) =: w^*(m)$$

which is independent of x. Now, $c(m) = w^*(m)$ is positive. It is clear that $w^*(m) = 0$ because $v^*_{j,\mu}(x,m) - w^*(m)$ satisfies the comparison property. i.e $v^*_{j,\mu}(x,m) - w^*(m) \geq v^*_{j,\mu}(m)$ from which we deduce $w^*(m) \leq 0$ (note that by unichain argument $v^*_{j,\mu}(x,m)$ does not depend on x). Now, consider the term $e^{\frac{1}{\mu}\sum_{t=0}^{T} r(x_t,a_t,m_t)}$. By positivity of $r(.)$ and monotonicity, one has

$$e^{\frac{1}{\mu}\sum_{t=0}^{T} r(x_t,a_t,m_t)} \longrightarrow e^{\frac{1}{\mu}\sum_{t=0}^{+\infty} r(x_t,a_t,m_t)}$$

when T goes to infinity. Then,

$$e^{\frac{1}{\mu}\sum_{t=0}^{T} r(x_t,a_t,m_t)}g(v^*_{j,\mu}(x_{T+1},m_{T+1})) \longrightarrow$$
$$e^{\frac{1}{\mu}\sum_{t=0}^{+\infty} r(x_t,a_t,m_t)}g(w^*(m)).$$

Thus,

$$\lim_{T}\mathbb{E}_{\pi}\left(e^{\frac{1}{\mu}\sum_{t=0}^{T} r(x_t,a_t,m_t)}g(v^*_{j,\mu}(x_{T+1})) \mid x,m\right) =$$
$$\mathbb{E}_{\pi,x,m}\left(e^{\frac{1}{\mu}\sum_{t=0}^{+\infty} r(x_t,a_t,m_t)} \mid x,m\right)$$

and $v^*_{j,\mu}(x,m) = F_{j,\mu}(\pi,x,m)$ which gives the optimality π in response to m. This completes the proof.