2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

Mean field stochastic games: convergence, Q/H-learning and optimality

Hamidou Tembine

Abstract— **We consider a class of stochastic games with finite number of resource states, individual states and actions per states. At each stage, a random set of players interact. The states and the actions of all the interacting players determine together the instantaneous payoffs and the transitions to the next states. We study the convergence of the stochastic game with variable set of interacting players when the total number of possible players grow without bound. We provide sufficient conditions for mean field convergence. We characterize the mean field payoff optimality by solutions of a coupled system of backward-forward equations. The limiting games are equivalent to discrete time anonymous sequential population games or to differential population games. Using multidimensional diffusion processes, a general mean field convergence to coupled stochastic differential equation is given. Finally, the computation of mean field equilibria is addressed using Q/H learning.**

I. Introduction

Dynamic Game Theory deals with sequential situations of several decision makers (often called players) where the objective for each one of the players may be a function of not only its own preference and decision but also of decisions of other players.

Dynamic games allow to model sequential decision making, time-varying interaction, uncertainty and randomness of interaction by the players. They allow to model situations in which the parameters defining the games vary in time and players can adapt their strategies (or policies) according the evolution of the environment. At any given time, each player takes a decision (also called an action) according to some strategy. A (behavioral) strategy of a player is a collection of history-dependent maps that tell at each time the choice (which can be probabilistic) of that player. The vector of actions chosen by players at a given time may determine not only the payoff for each player at that time; it can also determine the state evolution. A particular class of dynamic games widely studied in the literature is the class of *stochastic games*. Those are dynamic games with probabilistic state transitions (stochastic state evolution) controlled by one or more players. The discrete time state evolution is often modeled as interactive Markov decision processes while the continuous time state evolution is referred to stochastic differential games. Discounted stochastic games have been introduced in [6]. Stochastic games and interactive Markov decision processes are widely used for modeling sequential decision-making problems that arise in engineering, computer science, operations research, and social sciences. However, it is well known that many real-world problems modeled by stochastic games have huge state and/or action spaces, leading to the well-known *curse of dimensionality* that makes solution of the resulting models intractable. In addition, if the size of the system grows without bound, the number of parameters: states, actions, transitions explode exponentially.

In this paper we develop a mean field limit for stochastic games with variable number of interacting players. Centralized and decentralized mean field solutions are obtained by identifying a consistency relationship between the individual-state-mass interaction such that in the population limit each individual optimally responds to the mass effect and these individual strategies also collectively produce the same mass effect presumed initially. This leads to a coupled system forward/backward optimality equations (partial differential equation or difference equations).

Short overview of mean field stochastic games in discrete time Mean field interactions with spatially distributed players and types can be described as a sequence of dynamic games. Since the population profile involves many players for each type or class, a common approach is to ignore 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, also known as non-atomic or population games have been studied by addressed by von Neumann (1944), Nash (1951). In the context of transportation networks, Wardrop (1952) have studied population games in a deterministic and stationary setting of indistinguishable players.

Discrete time mean field games with continuum of players have been studied in [4] under the name *Anonymous sequential games*. Classes of mean field approximations and oblivious equilibria have been studied in [1], [13]. Transition from discrete to continuous time mean field stochastic games are considered in [8], [10]. Bergin and Bernhardt [2] showed how *stochastic mean field limit* can be introduced into the model (so the mean field limit evolves stochastically). In [11], we have applied noisy mean field limit in malware propagation in opportunistic networks.

What is new in the mean field game approach? In the mean field Markov game modeling in discrete time, there must be an equation to express the dynamic optimization problem of each player. Usually this involves one equation for each player. If players are classified together by similar player classes, there is one equation per class. This equation is generally a Bellman-Shapley equation, since a large proportion of optimization problems fall within the

H. Tembine is with Ecole Supérieure d'Electricité, SUPELEC, France. E-mail: tembine@ieee.org

978-1-4577-0079-8/11/$26.00 ©2011 AACC

framework of dynamic programming. Hence, the Bellman-Shapley equations will be used to compute optimal behavioral strategies. An equation is also needed to express the class' behavior, the mean field behavior of each class. The dynamics of the distribution is governed by a Kolmogorov forward equation. In this Kolmogorov forward equation, the optimal behaviors of the players occur as data, since it is the infinite collection of individual behaviors that is aggregated and constitutes collective behavior by consistency. Thus, the modeling of the behavior of a group of players (sub-population) naturally leads to a BS-K (Bellman-Shapley and Kolmogorov) system of equations. The discrete BS-K have been studied by Jovanovic & Rosenthal in the eighty's. The novelty in their study is that the mean field games formalism involves *the density of players on the state space can enter in the Bellman-Shapley equation*. Thus, the mean field equilibrium is defined by an BS-K system in which the Bellman-Shapley equations are doubly coupled: individual behaviors are given for the Kolmogorov forward equation and, at the same time, the distribution of players in the state space enters in the Bellman equation which is completely innovative. This means that players can incorporate into their preferences the density of states/actions of other players at the anticipated equilibrium. Therefore each player can construct his strategy by taking account of the anticipated distribution of strategies and of the actions of other players. Under suitable conditions, this fixed-point of behaviors, the mean field equilibria can be defined by moving to the limit on the number of players in the class of Markov games in discrete time (or difference games) that are asymptotically invariant (in law) by permutation within the same class of players called *Asymptotic Indistinguishability Per Class Property*.

Our contribution can be summarize as follows. We provide mean field convergence results for a class of mean field stochastic games with multiple classes of players. We characterize the mean field limit as a solution of *Kolmogorov forward equations*. Then, we formulate individual dynamic optimization problem in which each player optimizes its expected long-term payoff under individual stochastic dynamics and mean field limit. The mean field solutions are obtained by identifying a consistency relationship between the individual-state-mass interaction such that in the population limit each individual optimally responds to the mass effect and these individual strategies also collectively produce the same mass effect presumed initially. This leads to a coupled system *forward-backward equations*. The existence of solutions, the computation of the solutions as well as Q/H-learning aspects are discussed.

The remainder of the paper is structured as follows. In the next section we present the model description. We then focus on mean field convergence. In section II sufficiency conditions for convergence to deterministic mean field limit are provided. When these conditions fail, some possible extension are given in section III in which the convergence to stochastic mean field limit is presented. Due to the space limitation, all the proofs are omitted.

II. DISCRETE TIME MEAN FIELD MODEL

In this section, we describe the controlled mean field interaction model. Time $t \in \mathbb{N}$ is discrete. There is a set of resources those states are represented by $S^n(t) \in \mathcal{S}$ (finite). There are n players ($n \geq 2$). For every player j, $\mathcal{X}$ is its own state space. An individual state has two components as follows: the type of the player and the internal state. The type is a constant during the game. The state of player j at time t is denoted by $X_j^n(t) = (\theta_j, Y_j^n(t))$ where θ_j is the type. The set of possible states $\mathcal{X}_j = \{1, 2, ..., \Theta\} \times \mathcal{Y}_j$ is finite. $\mathcal{Y}_j$ may include other parameters, such as, space location, current direction and so on. The individual state of player j at time t is denoted by $X_j^n(t)$. For every player j, $\tilde{\mathcal{A}}_j$ is the set of actions of that player. $\mathcal{A}_j : \mathcal{S} \times \mathcal{X}_j \longrightarrow 2^{\tilde{\mathcal{A}}_j}$ is a set-valued map (correspondence) that assigns to each state $(s, x_j) \in \mathcal{S} \times \mathcal{X}_j$ the set of actions $\mathcal{A}_j(s, x)$ that are available to player j. We assume that the set $\mathcal{A}_j(s, x)$ depends only on the type θ_j and value of the state x_j. The action of player j at time t is $A_j^n(t)$. The *global state* of the system at time t is $(S^n(t), X^n(t)) = (S(t), X_1^n(t), ..., X_n^n(t))$. Denote by $A^n(t) = (A_1^n(t), \ldots, A_n^n(t))$ the action profile at time t. The system $(S^n(t), X^n(t))$ 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 player coupled not only via their instantaneous payoff function by $r^n(S^n(t), X^n(t), A^n(t))$ but also via the state evolution $X^n(t)$ i.e the evolution of $X_j^n(t)$ depends on the states and the actions of the other players.

For $u \in \mathcal{U}$, define $M^n[u](t)$ to be the current population profile i.e

$$M_x^n[u](t) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{X_j^n(t)=x\}}. \tag{1}$$

At each time t, $M^n[u](t)$ is in the finite set $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}^{|\mathcal{X}|}$, and $M_x^n[u](t)$ is the fraction of players who belong to population of individual state x. For a subset $X_1 \subseteq \mathcal{X}$, define $M^n[u](t)(X_1') := \frac{1}{n} \sum_{j=1}^n \delta_{\{X_j^n[u](t) \in X_1'\}}$.

Note that $X_j^n[u](t)$ (and hence and $M^n[u](t)$) depends implicitly on the decision process. To simplify the notations, we write $X_j^n(t)$ (resp. $M^n(t)$) to denote the state process driven by u (resp. the mean field process driven by u).

Similarly, we associate the process $U_a^n(t) = \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{\{A_j^n(t)=a\}}$ to the fraction of players per action.

Strategies and random set of interacting players: At time slot t, an ordered list $\mathcal{B}_t^n$, of players in $\{1, 2, \ldots, n\}$, without repetition, is selected randomly as follows. First we draw a random number of players k_t such that $\mathbb{P}(|\mathcal{B}_t^n| = k \mid M^n(t) = m) =: J_k^n(m)$ where the distribution $J_k^n(m)$ is given for any n, $m \in \{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}^{|\mathcal{X}|}$. Second, we set $\mathcal{B}_t^n$ to an ordered list of k_t players drawn uniformly at random among the $n(n-1)...(n-k_t+1)$ possible ones.

Each player such that $j \in \mathcal{B}_t^n$ takes part in a one-shot interaction at time t, as follows. First, each selected player $j \in \mathcal{B}_t^n$ chooses an action $a_{j,t} \in \mathcal{A}(s, x_j)$ with probability $u(a_j \mid s, x_j, n, t)$ where (s, x_j) is the current state of that player and the state of the resource. The stochastic array

u can be interpreted as the strategy profile of the hull population. We write u^n_t to denote a generic Markov strategy at time t.

Denoting the current set of interacting players $\mathcal{B}^n_t = \{j_1, \ldots, j_k\}$. Given the actions $a_{j_1}, ..., a_{j_k}$ drawn by the k players, we draw a new set of individual states $(x'_{j_1}, ..., x'_{j_k})$ and resource state s' with probability $L^n_{s;s'}(k, m, a)$, where a is the vector of the selected actions by the interacting players.

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 once the sequence of strategy is given. Denote by $w^n_{s,s'}(u, m)$ be the marginal transition probability between the resource states. Given any Markov strategy and any vector m of $\Delta(\mathcal{X})$, the resource state generates an irreducible Markov decision process with limiting invariant measure $w_s(u, m)$. Then, we can simplify the analysis by fixing the resource state $S(t) = s$ without losing generality.

A. Kernel definitions

We provide a general convergence result of the mean field to a stochastic differential equation and a martingale problem for the law of process $M^n(t)$. Let $\mathcal{F}^n_t = \sigma(S(t'), X^n(t'), A^n(t'),\ t' \leq t)$ 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 $h_t = (S(0), X^n(0), A^n(0), \ldots, S(t) = s, X^n(t), A^n(t)) \in \mathcal{F}^n_t$. $X^n(t+1)$ evolves according to the transition probability

$$L^n(x'; x, u, s) = \mathbb{P}\left(X^n(t+1) = x' \mid \mathcal{F}^n_t\right).$$

The term $L^n(x'; x, u, s)$ is the transition kernel on $\mathcal{X}^n$ under the strategy U^n. Let $x^n = (x^n_1, \ldots, x^n_n)$ such that $\frac{1}{n}\sum_{j=1}^n \delta_{x^n_j} = m$ and define

$$\mathcal{L}^n(m'; m, u, s) = \sum_{\substack{(x'_1, \ldots, x'_n) \\ \frac{1}{n}\sum_{j=1}^n \delta_{x'_j} = m'}} L^n(x'; x, u, s).$$

The system evolves according to the kernel

$$\begin{aligned} &\mathcal{L}^n(m'; m, u, s) \\ := \mathbb{P}(M^n(t+1) = m' \mid &M^n(t) = m, U^n(t) = u, S(t) = s) \\ &= \mathbb{P}(M^n(t+1) = m' \mid \tilde{h}_t) \end{aligned}$$

where $\tilde{h}_t = (S(t'), X^n(t'), A^n(t'),\ t' \leq t, S(t) = s, X^n(t) = x^n)$, such that $\frac{1}{n}\sum_{j=1}^n \delta_{x^n_j} = m$. The term $\mathcal{L}^n(m'; m, u, s)$ corresponds to the projected kernel of L^n.

III. Mean field convergence

In this section, we present two main mean field convergence results.

A. Deterministic mean field limit

Proposition 1: Let $\mathcal{M}^d_n = \{m \mid nm \in \mathbb{N}^d\}$. Suppose that

A0: For every s, the function $w_s(u, m)$ is continuously differentiable in m and u.

A1: $\exists\ 0 < \delta_n, \epsilon_n \searrow 0$, and a continuously differentiable function $f:\ \mathbb{R}^d \times \mathcal{U} \times \mathcal{S} \longrightarrow \mathbb{R}^d$ such that

$$\lim_n \sup_{u \in \mathcal{U}} \sup_{\|m\| \leq 1} \parallel \frac{f^n(m, u, s)}{\delta_n} - f(m, u, s) \parallel = 0,$$

where $x \in \mathcal{X}$ and $f^n_x(m, u, s) =$

$$\int_{m' \in \mathcal{M}^d_n} 1\!\!1_{\|m'-m\| \leq 2}(m'_x - m_x)\mathcal{L}^n(dm'; m, u, s),$$

A2:

$$\sup_n \sup_{u \in \mathcal{U}} \frac{1}{\delta_n} \int_{m' \in \mathcal{M}^d_n} \parallel m' - m \parallel \mathcal{L}^n(dm'; m, u, s) < +\infty$$

A3: $\lim_n\ \sup_{u \in \mathcal{U}} \frac{1}{\delta_n} \int_{m' \in \mathcal{M}^d_n} 1\!\!1_{\|m'-m\| > \epsilon_n} \parallel m' - m \parallel \mathcal{L}^n(dm'; m, u, s) = 0,$

A4: $M^n(0) = m^n_0$ converges to $m_0 \in \Delta(\mathcal{X})$.

Then, for all $\epsilon > 0, T < +\infty$,

$$\lim_n \mathbb{P}\left(\sup_{t \in [0,T]} \parallel \tilde{M}^n(\frac{t}{\delta_n}) - m(t, u, m_0) \parallel > \epsilon\right) = 0,$$

where $\tilde{M}^n_t$ is the interpolated process from M^n_t, $m(t, u, m_0)$ is the unique solution of the ordinary differential equation $\dot{m}_t = \tilde{f}(u_t, m_t)$ starting from $m_0 \in \Delta(\mathcal{X})$ where

$$\tilde{f}(u_t, m_t) := \sum_{s \in \mathcal{S}} w_s(u_t, m_t) f(u_t, m_t, s).$$

The assumption A1 demands that as n grows large, the expected changes per time unit $\frac{f^n}{\delta_n}$ converge uniformly to a Lipschitz continuous vector field f. Lipschitz continuity of f ensures the existence and uniqueness of solutions of the mean field game dynamics $\dot{m}_t = f(u_t, m_t), m(0) = m_0$ The assumption A2 requires that the expected absolute changes per time unit is bounded. The assumption A3 demands that jumps larger than ϵ_n make vanishing contributions to the motion of the processes, where ϵ_n is a sequence of constants that converges to zero. A0 and A4 are respectively regularity assumptions and initialization conditions.

Consequently, under the vanishing scaling assumptions δ_n, ϵ_n and the hypothesis A0-A4, one has a deterministic approximation of the random process M^n and the deterministic trajectory is described by the ODE.

As we can see some of the assumptions in the above proposition may not be satisfied in wide range of applications in large population. The assumptions are satisfied when the second moment of number of players that change their individual states in one time slot are bounded in expectation. However, when there are simultaneous and many local interactions as large population games, the second moment may not be finite when the size of the population goes to infinity. Then, a natural question is to ask is: what will happens if the second moment condition is not satisfied?

In the next section, we will partially answer to this question by proving a mean field convergence to controlled stochastic differential equation called *noisy mean field limit* in [7].

B. Stochastic mean field limit

Below we provide sufficient conditions on the transition kernels $\mathcal{L}^n$ to get a weak convergence of the process M_t^n under the strategy $U^n(t)$.

B0: For every $s \in \mathcal{S}$, $w_s(u,m)$ is continuously differentiable in m and u.

B1: There exists $\delta_n \searrow 0$ and continuous mapping $a:\ \mathbb{R}^d \times \mathcal{U} \times \mathcal{S} \longrightarrow \mathbb{R}^{d\times d}$ such that $(x,x',s) \in \mathcal{X}^2 \times \mathcal{S}$,

$$\lim_n \sup_{u\in\mathcal{U}} \sup_{\|m\|\leq 1} \parallel \frac{a^n(m,u,s)}{\delta_n} - a(m,u,s) \parallel = 0,$$

where $(x,x',s) \in \mathcal{X}^2 \times \mathcal{S}$, $a^n_{x,x'}(m,u,s) =$

$$\int_{m'} \mathbb{1}_{\|m'-m\|\leq 2}(m'_x - m_x)(m'_{x'} - m_{x'})\mathcal{L}^n(dm';m,u,s),$$

and the third moment is finite. Denote by

$$\tilde{a}_{x,x'}(m,u) = \sum_{s\in\mathcal{S}} w_s(m,u) a_{x,x'}(m,u,s)$$

B2: There exists a continuous mapping $f:\ \mathbb{R}^d \times \mathcal{U} \times \mathcal{S} \longrightarrow \mathbb{R}^d$ such that $\forall s \in \mathcal{S}$,

$$\lim_n \sup_{u\in\mathcal{U}} \sup_{\|m\|\leq 1} \parallel \frac{f^n(m,u,s)}{\delta_n} - f(m,u,s) \parallel = 0,$$

B3: For all $\epsilon > 0$; $\forall s \in \mathcal{S}$,

$$\lim_n \sup_{u\in\mathcal{U}} \frac{1}{\delta_n} \int_{m'\in\mathbb{R}^d} \mathbb{1}_{\|m'-m\|>\epsilon} \mathcal{L}^n(dm';m,u,s) = 0,$$

B3’: $\forall s \in \mathcal{S}$,

$$\sup_{u\in\mathcal{U}} \sup_{m\in\mathcal{R}^d} \sup_{n\geq 1} \left[\parallel \frac{a^n(m,u,s)}{\delta_n} \parallel + \parallel \frac{f^n(m,u,s)}{\delta_n} \parallel \right] < \infty,$$

Proposition 2: Assume $B0 - B3$. Then, for any test function ϕ, the generator $\frac{1}{\delta_n}\mathcal{L}^n\phi(m,u) \longrightarrow \mathcal{L}\phi(m,u)$ for any m,u where $\mathcal{L}\phi(m,u) = \sum_x \tilde{f}_x(m,u)\frac{\partial}{\partial m_x}\phi(m,u) + \frac{1}{2}\sum_{x,x'} \tilde{a}_{x,x'}(m,u)\frac{\partial^2}{\partial m_x \partial m_{x'}}\phi(m,u)$.

Moreover, if the function $\tilde{a}(.,.)$ and $\tilde{f}(.,.)$ have the property that for each $(m,u) \in \mathbb{R}^d \times \mathcal{U}$, the martingale problem for a and f has exactly one solution $\pi_{m,u}$ starting from m. Then $\pi_{n,m,u} \longrightarrow \pi_{m,u}$ as $\delta_n \searrow 0$ uniformly in m for any strategy $u \in \mathcal{U}$ where $\pi_{n,m,u}$ is the law of interpolated process from $M^n(t)$. In addition, if $B3'$ holds then the martingale problem has a unique solution.

This result provides a mean field convergence to a solution of stochastic differential equation with drift f and diffusion term a which is reported in the following corollary:

Corollary 1: Suppose that $M_0^n \longrightarrow \mu_0$ in law where μ_0 is a probability measure. Under B0-B3’, the process M_t^n converges in law to a solution of the stochastic differential equation (SDE) given by

$$d\tilde{m}_t = \tilde{f}(u_t,\tilde{m}_t)dt + \tilde{\sigma}(u_t,\tilde{m}_t)d\mathbb{B}_t$$

where $\tilde{\sigma}\tilde{\sigma}^t = \tilde{a}$, and $\mathbb{B}$ is a standard Brownian motion (a Wiener process).

This result follows from Proposition 2 and the convergence of $\frac{1}{\delta_n}\mathcal{L}^n\phi(m,u,s) \longrightarrow \mathcal{L}\phi(m,u,s)$ using the tightness properties of the processes.

C. Connection to propagation of chaos

Under the above assumptions, we have shown the convergence of the process $(M^n(t))_{t\in[0,T]}$ for any $T \in \mathbb{R}$. Since our mean field stochastic games model satisfies the invariance in law by any permutation with players index within the same type if the control u satisfies this property, one can use the *asymptotic indistinguishability per class* or indistinguishability per class to establish a propagation of chaos. Let $X_j^n = (X_j^n(t))_{t\geq 0}$. The process $\bar{M}^n = \frac{1}{n}\sum_{j=1}^n \delta_{X_j^n}$ converges in law to the process $\tilde{m}$ with law μ and for any k, any measurable and bounded functions $\phi_1,\ldots,\phi_k$

$$\lim_n \mathbb{E}\left(\prod_{j=1}^k \phi_j(X_j^n)\right) = \prod_{j=1}^k \left(\int \phi_j d\mu\right) \tag{2}$$

Note that the propagation of chaos property holds if the ODE (resp. the SDE) has a unique global attractor m^*. However the propagation of chaos property may not hold in stationary regime, see [9].

IV. Mean field optimality

A. Vanishing step-size

From the above sections, we know that under the assumptions A0-A4, the mean field limit is deterministic and it is the unique solution of the ODE

$$\dot{m}_t = f(u_t,m_t),\ m(0) = m_0 \in \Delta(\mathcal{X}). \tag{3}$$

Consider the long-term payoff $F_T(u,m) = \int_t^T r(u_{t'},m_{t'})\ dt' + g(m_T)$ subject to the constraint of (3). Let v be the optimal value i.e $v = \sup_u F_T(u,m)$.

The mean field optimality[1] for horizon T is given by

$$\left\{\begin{array}{c} v(T,m) = g(m) \\ -\frac{\partial}{\partial t}v(t,m_t) = \sup_{u_t\in\mathcal{U}}\left\{r(m_t,u_t)\right. \\ \left. +\sum_{x\in\mathcal{X}}\tilde{f}_x(m_t,u_t)\frac{\partial}{\partial m_x}v(t,m)\right\} \\ m_t = m_0 + \int_0^t \tilde{f}(m_{t'},u^*_{t'})dt',\ t' > 0 \\ m_0 = m. \end{array}\right.$$

which is a backward-forward system of differential equations.

Similarly, when considering the assumptions B0-B3’, the mean field limit is stochastic and given by the stochastic differential equation $d\tilde{m}_t = \tilde{f}(u_t,\tilde{m}_t)dt + \tilde{\sigma}(u_t,\tilde{m}_t)d\mathbb{B}_t$, $\tilde{m}(0) = \tilde{m}_0 \in \Delta(\mathcal{X})$. For the expected long-term payoff with horizon T, $\tilde{F}_T(t,u,m) = \mathbb{E}\left(\int_t^T r(u_{t'},m_{t'})\ dt' + g(m(T))\right)$. Under regularity of the above functions, we use Itô’s formula for the payoff evolution for a fixed horizon T and a fixed u and we apply dynamic programming principle under noise to get the following mean field optimality:

[1] Note that the term “mean field optimality” does not refer the optimality of the strategy. Here we restrict our attention to mean field responses which are not necessarily optimal in the finite regime.

$$\begin{cases} \tilde{v}(T,m) = g(m), \\ -\frac{\partial}{\partial t}\tilde{v}(t,m_t) = \sup_{u_t \in \mathcal{U}} \{ r(m_t,u_t) \\ \quad + \sum_{x\in\mathcal{X}} \tilde{f}_x(m_t,u_t)\frac{\partial}{\partial m_x}\tilde{v}(t,m_t) \\ \quad +\frac{1}{2}\sum_{(x,x')\in\mathcal{X}^2} \tilde{a}_{x,x'}(m_t,u_t)\frac{\partial^2}{\partial m_x \partial m_{x'}}\tilde{v}(t,m_t) \Big\} \\ \partial_t m_t + div\left(\tilde{f}(m_t,u_t^*)m_t\right) = \frac{1}{2}\sum_{x,x'}\partial^2_{xx'}\left(\tilde{a}_{xx'}(m_t,u_t^*)m_t\right) \\ m_0 = m. \end{cases}$$

The result follows by combining Itô-Dynkin's formula and Hamilton-Jacobi-Bellman-Fleming's stochastic optimality equation which is obtaining by using the dynamic programming principle associated to the maximization problem.

The first equation of the system is a Hamilton-Jacobi-Bellman-Fleming (HJBF) backward equation and the second equation of the last system is called Fokker-Planck-Kolmogorov (FPK) forward equation.

Now, considering any generic player j which optimizes its own long-term payoff, we define the best response to $\{m_t\}_t$ denoted $BR(m)$ as the set of strategy $\{u_t\}_t$ that are maximizers of the individual payoff of any generic player when the mean field limit follows $\{m_t\}_t$.

We say that the pair (u^*, m^*) such that $(u^*_t, m^*_t) \in \mathcal{U} \times \Delta(\mathcal{X})$ is a mean field equilibrium if $u^* \in BR(m^*)$ and m^* is generated by u^*.

B. Non-vanishing step-size

In this subsection, we analyze the case where the step-size is not vanishing when n goes to infinity. In that case, the mean field limit is in discrete time and driven by the probability transition $L_t(u,m)$. Given a initial population profile m_0 and a terminal payoff g, the sequence of population profile $\{m_t\}_t$ is driven by the transition probabilities $\{L_{t,x,x'}(u_t,m_t)\}_t$.

$$m_{t+1}(x) = \sum_{x'\in\mathcal{X}} m_t(x')L_{t,x,x'}(u_t,m_t). \tag{4}$$

where $L_{t,x,x'}(u,m) = \sum_{k\geq 0}\sum_s w_s(u,m)L_{t,s,x,x'}(u,m;k)J_k(m)$, $L_{t,s,x,x'}(u,m;k)$ is the limiting probability transition from x to x' when the resource state is s and the number of interacting players is k. Combining with the Bellman-Shapley optimality criterion, one gets the following system in the finite horizon case:

$$\begin{cases} v_t(s,x) = \max_u \{ r(s,x,u,m_t) \\ + \sum_{s',x'} L(s',x'|s,x,u,m_t)v_{t+1}(s',x') \Big\} \\ m_{t+1}(x) = \sum_{x'\in\mathcal{X}} m_t(x')L_{t,x',x}(u_t,m_t) \end{cases}$$

Proposition 3: Under the above assumptions, the finite horizon (resp. the discounted) mean field stochastic game has at least one mean field equilibrium.

Similarly, backward-forward system of mean field optimality can be derived for the discounted payoff.

The proof of this result follows from Jovanovic & Rosenthal (1988). Of course, we need to check the assumptions: non-emptiness, compactness of the actions spaces, continuity and boundedness of the payoff functions. The idea of the proof is to construct a correspondence (set-valued mapping) in the space of measure "m" that satisfied (i) non-emptiness, closed-valued, convex-valued and upper-semi-continuity, (ii) compatibility with the intersection of the measures generated by the Kolmogorov forward equations. Then, we use Fan-Glicksberg-Debreu (1952) fixed-point theory to conclude.

Note that the above results do not say that the mean field equilibria are approximated equilibria for large n. In order to have approximated equilibria one may need additional assumptions on the payoff functions r^n and its relation to r. The same remark holds for centralized case: in general for a given approximating first then optimizing is different than optimizing first and approximating the optimizers in second step. In our situation on may need a specific structure of $\frac{1}{n}\sum_{j=1}^n r^n(X_j^n(t), X_{-j}^n(t), A^n(t))$.

V. Discussion on Mean field Q/H learning

In this section we discuss on mean field Q/H learning for stochastic games.

A. Mean field Q-learning

Q-learning is a technique used to compute an optimal policy for a controlled Markov chain based on observations of the system controlled using a non-optimal policy. Many interesting results have been obtained for models with finite state and action space.

While an optimal strategy can, in principle, be obtained by the methods of dynamic programming, policy iteration, and value iteration, such computations are often prohibitively time-consuming. In particular, if the size of the state space grows exponentially with the number of state variables, a phenomenon referred to by Bellman as the *curse of dimensionality*. Similarly, the size of the action space can also lead to computational intractability. Following the idea of Q-learning [12], we define the Q-value $Q(s,x_j,a_j,m_t) = r_j(s,x_j,a,m_t) + \sum_{s',x'} L(s',x'_j \mid s,x_j,a_j,u_t,m_t)v(s',x'_j)$. Let $\sigma = (\sigma_t)_{t\geq 0}$, σ_t is a mapping that assigns to every finite history $h_{j,t}$ an element of $\Delta(\mathcal{A}_j(s_t,x_{j,t}))$, h_{jt} is a collection of history up to t that are available to player j. Every strategies σ , together with the initial state $s_0, x_{j,0}$ and the mean field m induces a probability distribution $\mathbb{P}_{s_0,x_{j,0},\sigma}$ over the space of infinite plays. We denote the corresponding expectation operator by $\mathbb{E}_{s_0,x_{j,0},\sigma}$. For the discounted payoff case with discount factor β_j, the payoff is

$$\begin{aligned} & F^n_{j,\beta_j}(s_0,x_{j,0},\sigma) \\ &= \mathbb{E}_{s_0,x_{j,0},\sigma}\left[\sum_{t\in\delta_n\mathbb{N}} \beta_j^t r_j^{\mathcal{B}_t^n}(s_t,x_t,a_t)\mathbb{1}_{\{j\in\mathcal{B}_t^n\}}\right] \end{aligned} \tag{5}$$

Assuming that $j \in \mathcal{B}^n(t), \forall t \geq 0$ one gets,

$$\begin{aligned} & Q^*(s,x_j,a_j,m) = r_j(s,x_j,a_j,m) \\ & +\beta_j \sum_{s',x'_j} L(s',x'_j \mid s,x_j,a_j,m)\left(\sup_{b_j\in\mathcal{A}_j(s',x'_j)} Q^*(s',x'_j,b_j,m')\right) \end{aligned}$$

One nice feature is to learn the Q-value without the knowledge of the transition probabilities. The iterative version is given by

$$Q_{t+1}(s,x_j,a_j,m) = Q_t(s,x_j,a_j,m) \\ +\lambda_t(s,x_j,a_j)\left[r_j(s,x_j,a_j,m)+ \\ \beta_j\left(\sup_{b_j\in\mathcal{A}_j(s',x'_j)} Q_t(s',x'_j,b_j,m')\right) - Q_t(s,x_j,a_j,m)\right] \quad (6)$$

where $\lambda_t(s,x_j,a_j) \geq 0$ is a learning rate function satisfying $\sum_t \lambda_t(s,x_j,a_j) = +\infty, \sum_t \lambda_t(s,x_j,a_j)^2 < +\infty$, which is a standard assumption in stochastic approximation.

B. Mean field H-learning

What does Q-learning have to do with the Hamilton-Jacobi-Bellman equation and the Pontryagin maximum principle?

The answers to this question have been examined in detail in [5] in a deterministic setting. The authors show that the Hamiltonian appearing in nonlinear control theory is essentially the same as the Q-function that is the object of interest in Q-learning. They established a close connection between Q-learning and differential dynamic programming [3] state space and general action space. This allows us to address the H-learning (learning under the Hamiltonian function) in mean field limit. The analogue of the Q-function is the H-function defined by

$$H(u,m,p,M) = r(u,m) + \langle p, \tilde{f}\rangle + \frac{1}{2}trace(\tilde{a}M)$$

where p is a $d-$diemensional vector and $M \in \mathbb{R}^{d\times d}$. Denote by W_l an approximation value at iteration l. The algorithm is described as follows:

- Initialize m_0
- At iteration l use m_l to compute the payoff $r(.,m_l)$.
- Use the Hamilton-Jacobi-Bellman-Fleming optimality to obtain W_l
- compute the optimal control u_l^* via the H-function (Pontryagin maximum principle)
- Use u_l^* to get m_l^* solution of the Fokker-Planck-Kolmogorov forward equation

VI. Conclusion

We have studied mean field stochastic games with random number of interacting players and established a mean field convergence to stochastic differential equations under suitable conditions. Combining Itô-Dynkin's formula with dynamic programming principle, we derived mean field optimality criterion characterizing *mean field equilibria*.

Acknowledgments

The author would like to thank three anonymous reviewers, the seminar audience at GERAD, UIUC and INRIA Hipercom and RAP for their interesting comments on a preliminary version of this paper.

References

[1] S. Adlakha, R. Johari, G. Weintraub, and A. Goldsmith. Oblivious equilibrium for large-scale stochastic games with unbounded costs. *Proceedings of the IEEE Conference on Decision and Control*, 2008.

[2] J. Bergin and D. Bernhardt. Anonymous sequential games with aggregate uncertainty. *Journal of Mathematical Economics*, 21:543–562, 1992.

[3] D. H. Jacobson and D. Q. Mayne. Differential dynamic programming. *American Elsevier Pub. Co., New York, NY*, 1970.

[4] B. Jovanovic and R. W. Rosenthal. Anonymous sequential games. *Journal of Mathematical Economics*, 17:77–87, 1988.

[5] P. Mehta and S. Meyn. Q-learning and pontryagin's minimum principle. *in IEEE Proc. CDC*, 2009.

[6] L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 39(10):1095–1100, 1953.

[7] H. Tembine. Mean field stochastic games: Simulation, dynamics and network applications. *Supelec*, 2010.

[8] H. Tembine. Population games in large-scale networks. *LAP, 250 pages*, 2010.

[9] H. Tembine. Hybrid mean field game dynamics in large populations. *American Control Conference, ACC*, 2011.

[10] 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.

[11] H. Tembine, P. Vilanova, and M. Debbah. Noisy mean field model for malware propagation in opportunistic networks. *Gamenets*, 2011.

[12] C.I.C.H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

[13] G. Y. Weintraub, L. Benkard, and B. Van Roy. Oblivious equilibrium: A mean field approximation for large-scale dynamic games. *Advances in Neural Information Processing Systems*, 18, 2005.