Hybrid Mean Field Learning in Large-Scale Dynamic Robust Games

Hamidou Tembine, Mohamad Assaad

Hybrid Mean Field Learning in Large-Scale Dynamic Robust Games

H. Tembine[a], **M. Assaad**[a]

[a]Supelec, Plateau Moulon, Gif-sur-Yvette, France
{hamidou.tembine, mohamad.assaad}@supelec.fr

Keywords: hybrid learning, robust games, dynamics

Extended Abstract

One of the objectives in distributed interacting multi-player systems is to enable a collection of different players to achieve a desirable objective. There are two overriding challenges to achieving this objective:

The first one is related to the complexity of finding optimal solution. A centralized algorithm may be prohibitively complex when there are large number of interacting players. This motivates the use of adaptive methods that enable players to self-organize into suitable, if not optimal, alternative solutions.

The second challenge is limited information. Players may have limited knowledge about the status of other players, except perhaps for a small subset of neighboring players. The limitations in term of information induce robust stochastic optimization, bounded rationality and inconsistent beliefs.

In this work, we investigate asymptotic pseudo-trajectories of large-scale dynamic robust games under various COmbined fully DIstributed PAyoff and Strategy Reinforcement Learning (CODIPAS-RL [4, 5]). We consider the following problem:

$$\forall j \in \mathcal{N},\ ROP_j : \sup_{\mathbf{x}_j \in \mathcal{X}_j} \mathbb{E}_{\mathbf{w}\sim\mu} U_j(\mathbf{w}, \mathbf{x}_j, \mathbf{x}_{-j}) \tag{1}$$

where $\mathcal{N}$ is the set of players, $\mathcal{X}_j$ is a subset of a finite dimensional space, $\mathbf{x}_{-j} = (\mathbf{x}_{j'})_{j'\neq j}$, $\mathbf{w}$ is a random variable with law μ.

The main issue here is that the mathematical structure of the payoff function $U_j : \mathcal{W} \times \prod_{j'\in\mathcal{N}} \mathcal{X}_{j'} \longrightarrow \mathbb{R}$ is not known by player j. We develop a combined distributed strategic learning in order to learn both expected payoff function as well as the associated optimal strategies. The dynamic game is played as follows:

Time is slotted. We start at some initial state $\mathbf{w}_0$. At each time slot t, the state $\mathbf{w}_t$ is drawn according to its distribution. Each player chooses an action according to some strategy. Each player is able to observe a measurement of its payoff. Based on the recent measurement, each player updates its strategy and its payoff estimations. The game goes to the next time slot.

Here, the term *dynamic* refers to the time/information dependence and the term *robust game* refers *game under uncertainty*. A strategy update of the player is mainly dictated by its learning scheme. We assume each player can choose among a finite set $\mathcal{L}$ of different CODIPAS-RLs. The hybrid learning has the following form:

$$\begin{cases} x_{j,t+1}(s_j) = \sum_{l\in\mathcal{L}} \mathbb{1}_{\{l_{j,t}=l\}} f^{(l)}_{j,s_j}(\lambda_{j,t}, \mathbf{x}_{j,t}, a_{j,t}, u_{j,t}, \hat{\mathbf{u}}_{j,t}) \\ \hat{u}_{j,t+1}(s_j) = \sum_{l\in\mathcal{L}} \mathbb{1}_{\{l_{j,t}=l\}} g^{(l)}_{j,s_j}(\mu_{j,t}, a_{j,t}, u_{j,t}, \hat{\mathbf{u}}_{j,t}) \\ s_j \in \mathcal{A}_j, j \in \mathcal{N},\ t \geq 0 \\ \mathbf{x}_{j,0} \in \mathcal{X}_j,\ \hat{u}_{j,0} \in \mathbb{R}^{|\mathcal{A}_j|} \end{cases}$$

where $u_{j,t}$ a noisy payoff observed by player j at time t, $\mathcal{A}_j$ is a set with the same size as the dimension of $\mathcal{X}_j$, $\mathcal{L}$ is a finite set of learning patterns, $l_{j,t}$ is the learning pattern chosen of player j at time t, $a_{j,t}$ is the action

chosen by player j at time t, s_j denotes a generic element of $\mathcal{A}_j$, the real-numbers $\lambda_{j,t}$ and $\mu_{j,t}$ are respectively strategy-learning rate and payoff-learning rate. The vector $\hat{u}_{j,t}$ is the estimated payoffs of player j corresponding to each component. For each $l \in \mathcal{L}$, the function $f^{(l)}$ is well-chosen in order to guarantee the almost sure forward invariance of $\prod_j \mathcal{X}_j$ by the stochastic process $\{\mathbf{x}_t\}_{t\geq 0}$.

Assuming that (i) the functions $U_j(., \mathbf{x})$ are integrable with respect to $\mu.$, (ii) $\mathcal{X}_j$ is convex, non-empty, compact, (iii) each component of the hybrid CODIPAS-RL learning scheme can be written in the form of Robbins-Monro [1] or Kiefer-Wolfowitz [2], we examine the long-run behavior the interacting system. The convergence, non-convergence, stability/instability properties of the resulting dynamics as well as their connections to different solution concepts (equilibria, global optima, satisfactory solutions, hierarchical solutions) are discussed. Those include Lyapunov expected robust games, expected robust pseudo-potential games, S-modular expected robust games, games with monotone expected payoffs, aggregative robust games, risk-sensitive robust pseudo-potential games.

A particular case of interest of our work is when the size of the system goes to infinity i.e. $|\mathcal{N}| \longrightarrow +\infty$. Under technical assumptions on indistinguishability per class [3] of learning property, we derive a Fokker-Planck-Kolmorogorov equation associated the hybrid CODIPAS-RL with large population of players called *mean field learning*. To prove this statement, we follow the following steps:

- Derivation of the asymptotic pseudo-trajectories of the stochastic processes $\mathbf{x}_{j,t}$.
- Identification of ordinary differential equations or Itô's SDE approximation via multiple time-scale stochastic approximations
- Time-scaling and classification of payoff functions in different classes.
- Derivation of the law of the empirical measure over the strategies and its associated interdependent systems of partial differential equations.

References

[1] Herbert Robbins and Sutton Monro, "A Stochastic Approximation Method", *Annals of Mathematical Statistics*, 22, 3 (September 1951), pp. 400-407.

[2] J. Kiefer and J. Wolfowitz, "Stochastic Estimation of the Maximum of a Regression Function", *Annals of Mathematical Statistics* 23, 3 (September 1952), pp. 462-466.

[3] H. Tembine, J. Y. Le Boudec, R. ElAzouzi, E. Altman,"Mean Field Asymptotic of Markov decision evolutionary games", *International IEEE Conference on Game Theory for Networks*, Gamenets 2009.

[4] H. Tembine, "Dynamic Robust Games in MIMO Systems", *IEEE Trans Syst Man Cybern B Cybern*, Aug. 2011 Volume: 41, Issue:4, 990 - 1002 ISSN: 1083-4419, 2011.

[5] H.Tembine, "Distributed strategic learning for wireless engineers", *Notes*, 440 pages, Supelec, 2010.