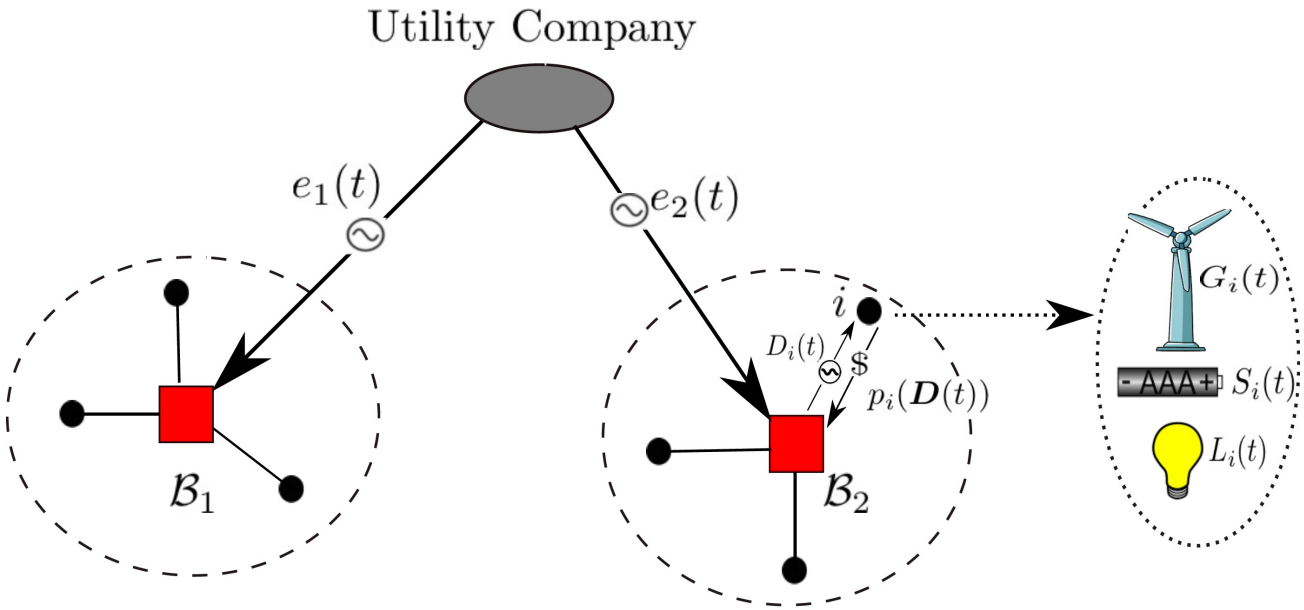
Learning Stationary Nash Policies in n-Player Stochastic Games with Independent Chains

Rasoul Etesami

Department of Industrial and Systems Engineering
Coordinated Science Lab
University of Illinois Urbana-Champaign, USA

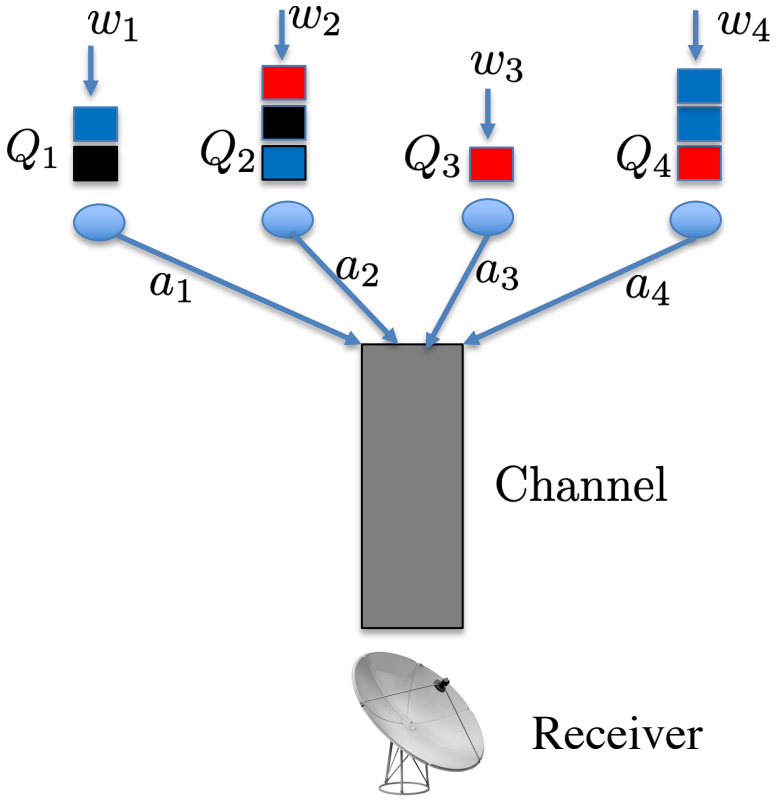
November 9, 2022
Linköping

Energy Management in Smart Grid



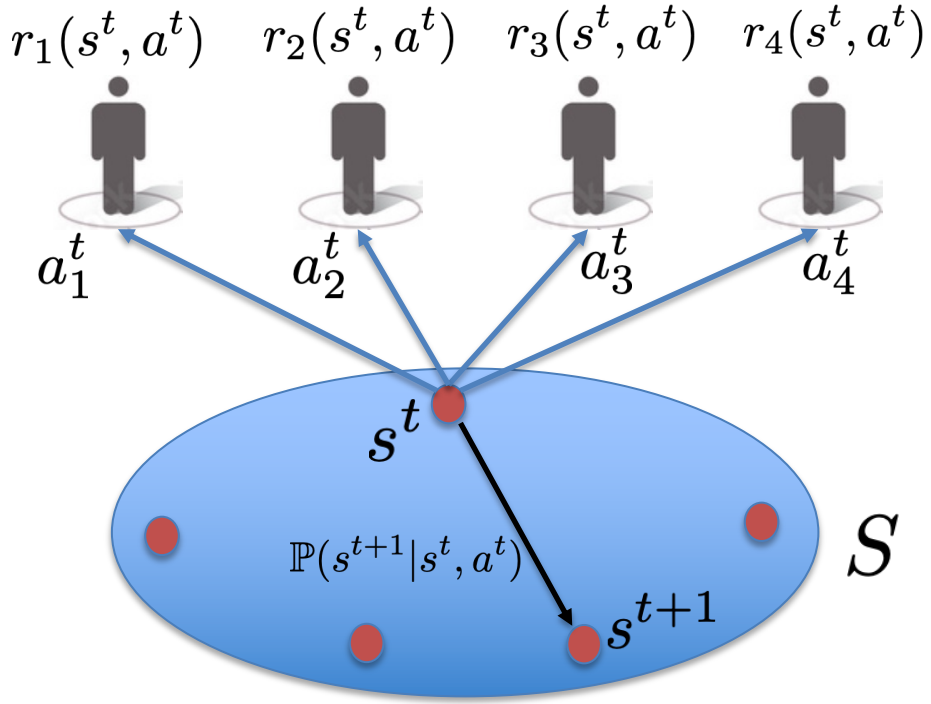
- Consider a utility company and n players, which can produce and consume energy.
- On day t, player i requests some units of energy from the utility company (action). The company sets the price as a function of aggregate demand and the available energy.
- As the harvested energy depends on the stochasticity of the weather conditions, the stored energy (state) of player i at the end of day t follows a stochastic process.
- The players want to adopt consumption policies to maximize their aggregate payoffs despite not being able to observe others' states/actions.

Multiagent Wireless Communication



- A set of users (players) sending messages to a common receiver over a wireless medium.
- At time t, player i looks at its state that is its buffer of size $Q_i(t)$ and decides whether to send a packet with some power.
- After that player i receives more packets based on some distribution from higher level.
- Players want to adopt policies to maximize their success transmission rates over time.

Stochastic Games



A policy for player i is a stationary policy if the probability of choosing action a_i^t at time t depends only on the current state s^t, and is independent of the time t, i.e., $\pi_i : S \to \Delta_{A_i}$

$$V_i(\pi_i, \pi_{-i}) = \mathbb{E}\Big[\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} r_i(s^t, a^t)\Big]$$

A policy profile $\pi^* = (\pi_1^*, \ldots, \pi_n^*)$ is called a Nash equilibrium if $V_i(\pi_i^*, \pi_i^*) \geq V_i(\pi_i, \pi_i^*)$ for any i and any policy π_i. It is called an ϵ-NE if $V_i(\pi_i^*, \pi_i^*) \geq V_i(\pi_i, \pi_i^*) - \epsilon$ for any i, π_i.

- **Stochastic games always admit a NE among stationary policies. (Shapley 1953)**

Related Work

[1] L. S. Shapley, “Stochastic games,” Proceedings of the National Academy of Sciences, vol. 39, no. 10, pp. 1095–1100, 1953.

[2] E. Altman, K. Avrachenkov, N. Bonneau, M. Debbah, R. El-Azouzi, and D. S. Menasche, “Constrained cost-coupled stochastic games with independent state processes,” Operations Research Letters, vol. 36, no. 2, pp. 160–164, 2008.

[3] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou, “The complexity of computing a Nash equilibrium,” SIAM Journal on Computing, vol. 39, no. 1, pp. 195–259, 2009.

[4] N. Cesa-Bianchi and G. Lugosi, Prediction, Learning, and Games. Cambridge University Press, 2006.

[5] E. Even-Dar, Y. Mansour, and U. Nadav, “On the convergence of regret minimization dynamics in concave games,” in Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing, 2009, pp. 523–532.

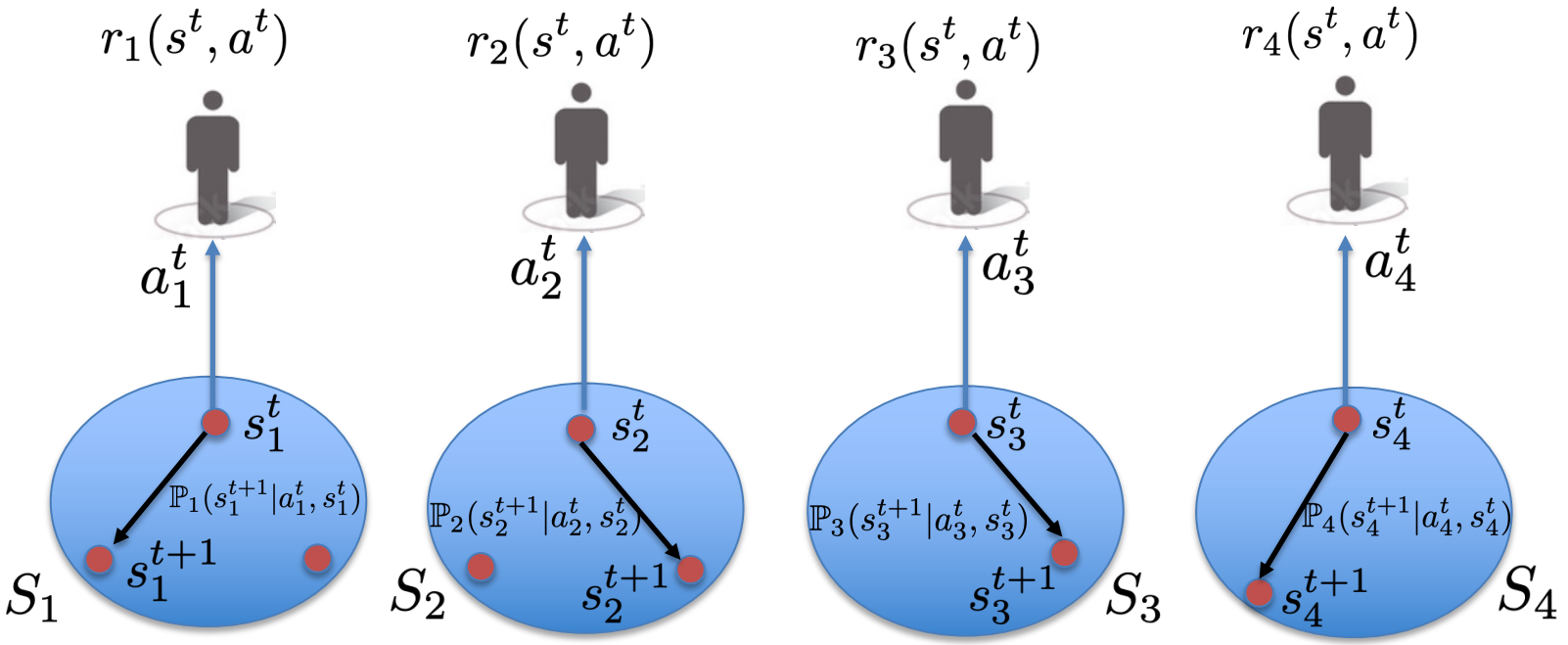
[6] P. Mertikopoulos and Z. Zhou, “Learning in games with continuous action sets and unknown payoff functions,” Mathematical Programming, vol. 173, no. 1, pp. 465–507, 2019.

[7] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, “On the theory of policy gradient methods: Optimality, approximation, and distribution shift,” Journal of Machine Learning Research, vol. 22, no. 98, pp. 1–76, 2021.

[8] C. Daskalakis, D. J. Foster, and N. Golowich, “Independent policy gradient methods for competitive reinforcement learning,” arXiv preprint arXiv:2101.04233, 2021.

[9] K. Zhang, Z. Yang, and T. Basar, “Multi-agent reinforcement learning: A selective overview of theories and algorithms,” Handbook of Reinforcement Learning and Control, Springer, pp. 321–384, 2021.

Stochastic Games with Independent Chains



A policy for player i is a stationary policy if the probability of choosing action a_i^t at time t depends only on its current state s_i^t, and is independent of the time t, i.e., $\pi_i : S_i \to \Delta_{A_i}$

$$V_i(\pi_i, \pi_{-i}) = \mathbb{E}\Big[\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} r_i(s^t, a^t)\Big]$$

- We assume that the joint transition probability $\mathbb{P}(s'|s,a)$ can be factored into independent components $\mathbb{P}(s'|s,a) = \prod_{i=1}^{n} \mathbb{P}_i(s_i'|s_i, a_i)$ where $\mathbb{P}_i(s_i'|s_i, a_i)$ is the transition model for i.

➢ **Stochastic games with independent chains always admit a stationary NE.**

A Dual Formulation

- Following of a stationary policy π_j by player j induces a Markov chain over S_j with corresponding stationary distribution ν_j, i.e., $\lim_{t\to\infty} \mathbb{P}(s_j^t = s_j) = \nu_j(s_j)$.

- For any player j, let us define ρ_j to be the occupancy probability measure that is induced over its state-action set $S_j \times A_j$ by following the stationary policy π_j, that is

$$\rho_j(s_j, a_j) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \mathbb{P}\{s_j^t = s_j, a_j^t = a_j\} = \nu_j(s_j)\pi_j(a_j|s_j).$$

- Therefore, we can write the payoff functions in terms of occupancy measures as

$$\begin{aligned} V_i(\pi_i, \pi_{-i}) &= \lim_{T\to\infty} \frac{1}{T} \sum_{s,a} \sum_{t=0}^{T} \mathbb{P}\{s^t = s, a^t = a\} r_i(s, a) \\ &= \sum_{s,a} \Big(\prod_{j=1}^{n} \nu_j(s_j)\pi_j(a_j|s_j) \Big) r_i(s, a) \\ &= \sum_{s,a} \prod_{j=1}^{n} \rho_j(s_j, a_j) r_i(s, a). \end{aligned}$$

A Dual Formulation

- Moreover, the set of feasible occupancy measures for player j can be written as the feasible points of the following polytope:

$$\mathcal{P}_i = \left\{ \rho_i \in \mathbb{R}_+^{|S_i \times A_i|} : \sum_{s_i, a_i} \left(P_i(s_i'|s_i, a_i) - \mathbb{I}_{\{s_i = s_i'\}} \right) \rho_i(s_i, a_i) = 0 \ \forall s_i', \ \sum_{s_i, a_i} \rho_i(s_i, a_i) = 1 \right\}.$$

Definition: The virtual game is an n-player static game, where the payoff and the action set for player i are given by $V_i(\rho) = \sum_{s,a} \prod_{j=1}^{n} \rho_j(s_j, a_j) r_i(s, a)$ and $\mathcal{P}_i$.

Lemma 1: (Altman 1999)

Given any MDP and any occupation measure ρ_i over its set of state-action, one can define a corresponding stationary policy

$$\pi_i(a_i|s_i) = \frac{\rho_i(s_i, a_i)}{\sum_{a_i' \in A_i} \rho_i(s_i, a_i')}, \quad \forall s_i \in S_i, a_i \in A_i,$$

such that following policy π_i in that MDP induces the same occupation measure as ρ_i over the state-action set $S_i \times A_i$.

Stochastic game with independent chains $\mathcal{G}$

$$V_i(\pi_i, \pi_{-i}) = \mathbb{E}\Big[\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} r_i(s^t, a^t)\Big]$$

Policy space $\pi_i : \mathcal{S}_i \to \Delta_{A_i}$

It can be played repeatedly

Dual formulation

Lemma 1

Static virtual game

$$V_i(\rho) = \sum_{s,a} \prod_{j\in[n]} \rho_j(s_j, a_j) r_i(s, a)$$

Occupation actions $\rho_i \in \mathcal{P}_i$

It can be played only once

Proposition

Finding a stationary NE for $\mathcal{G}$ without any assumption on the reward functions is PPAD-hard. → **No hope for scalable learning algorithms!**

- We restrict our attention to the cases where players' reward functions allow the existence of scalable learning algorithms.
- Using the equivalence between the stochastic game and the virtual game, we focus on finding a NE for the virtual game.
- We show how to repeatedly play over $\mathcal{G}$ and use the collected information in the virtual game to guide the learning dynamics to a NE.

Description of the Algorithm

- The algorithm proceeds in different episodes (batches), where each batch contains a random number of time instances.

- The occupation measure of player i at the beginning of batch k is denoted by ρ_i^k; during batch k, player i chooses actions according to the stationary policy π_i^k corresponding to ρ_i^k.

- A batch continues for a random number of time instances until each player i has visited all its states S_i at least once.

- Using the collected samples during batch k, player i constructs an (almost) unbiased estimator R_i^k for the gradient of its virtual payoff function $\nabla_{\rho_i} V_i(\rho)$.

- The estimator R_i^k is then used in a dual-averaging oracle with an appropriately chosen step-size/regularizer to obtain a new occupation measure ρ_i^{k+1}.

Algorithm 1 A Dual Averaging Algorithm for Player i

Input: Initial occupation measure $\rho_i^0 \in \mathcal{P}_i^{\delta_i}$, step-size sequence $\{\eta_i^k\}_{k=1}^{\infty}$, a fixed threshold d, initial dual score $Y^0 = \mathbf{0}$, and a strongly convex regularizer $h_i : \mathcal{P}_i^{\delta_i} \to \mathbb{R}$.

For $k = 1, 2, \ldots$, do the following:

- At the end of batch $k-1$, denoted by τ^k, compute

$$\pi_i^k(a_i|s_i) = \frac{\rho_i^k(a_i, s_i)}{\sum_{a_i' \in A_i} \rho_i^k(a_i', s_i)}, \quad \forall s_i \in S_i, a_i \in A_i,$$

  and keep playing according to this stationary policy π_i^k during the next batch k. Let $\tau_i^k \geq \tau^k + d$ be the first (random) time such that all states in S_i are visited during $[\tau^k + d, \tau_i^k]$. Batch k terminates at time $\tau^{k+1} = \max_i \tau_i^k$.

- Let $S_i' = S_i$, and $R_i^k \in \mathbb{R}_+^{|S_i||A_i|}$ be a random vector (initially set to zero), which is constructed sequentially during the sampling interval $[\tau^k + d, \tau^{k+1}]$ as follows:
  - **For** $t = \tau^k + d, \ldots, \tau^{k+1}$ and while $S_i' \neq \emptyset$, player i picks an action a_i^t according to $\pi_i^k(\cdot|s_i^t)$, and observes the payoff $r_i(s^t, a^t)$ and its next state s_i^{t+1}. If $s_i^t \in S_i'$, then update $S_i' = S_i' \setminus \{s_i^t\}$, and compute

$$R_i^k = R_i^k + \frac{r_i(s^t, a^t)}{\pi_i^k(a_i^t|s_i^t)} \, \mathbf{e}_{(s_i^t, a_i^t)},$$

    where $\mathbf{e}_{(s_i^t, a_i^t)}$ is the basis vector with all entries being zero except that the (s_i^t, a_i^t)-th entry is 1.
    End For

- In the end of batch k, compute the dual score $Y_i^{k+1} = Y_i^k + \eta_i^k R_i^k$, and update the occupation measure:

$$\rho_i^{k+1} = \operatorname*{argmax}_{\rho_i \in \mathcal{P}_i^{\delta_i}} \left\{ \langle \rho_i, Y_i^{k+1} \rangle - h_i(\rho_i) \right\}.$$

End For

Theorem 1

Given $\epsilon > 0$, suppose that each player i follows Algorithm 1 using a sequence of step-sizes that satisfy $\sum_{k=1}^{\infty} \eta_i^k = \infty$, and $\sum_{k=1}^{\infty} \left(\frac{\eta_i^k}{w_i^k}\right)^2 < \infty$, where $w_i^k = \sum_{\ell=1}^{k} \eta_i^\ell$. If with positive probability the sequence of occupation measures generated by Algorithm 1 converges to some point $\lim_{k\to\infty} \rho^k = \rho^*$, then the stationary policy corresponding to the limit point ρ^* is a stationary ϵ-NE.

How can we ensure convergence?

Assumption: A virtual game is called socially concave if i) there are positive constants $\lambda_i > 0$ such that $\sum_{i=1}^{n} \lambda_i V_i(\rho)$ is a concave function of ρ, and ii) for any player i and any fixed ρ_i, the payoff function $V_i(\rho_i, \rho_{-i})$ is a convex function of ρ_{-i}.

- Any 2-player game
- Linear resource allocation games
- Linear Cournot games
- TCP congestion control games
- …

Theorem 2

Assume that the virtual game is socially concave and the sequence of step-sizes η^ℓ satisfy $\sum_{\ell=1}^{\infty} \eta_i^\ell = \infty$ and $\sum_{\ell=1}^{\infty} (\eta_i^\ell)^2 < \infty$. Given $\epsilon > 0$, if all players follow Algorithm 1, the average occupations $\bar{\rho}_i^k = \sum_{\ell=1}^{k} \frac{\eta_i^\ell}{w_i^k} \rho_i^\ell$, where $w_i^k = \sum_{\ell=1}^{k} \eta_i^\ell$, will be an ϵ-NE almost surely, as $k \to \infty$.

Theorem 3

Let $\alpha \in (0, 1)$, and assume that each player follows Algorithm 1. Under the same assumptions as in Theorem 2, with probability at least $1 - \alpha$, the average occupancy $\bar{\rho}^k$ is an ϵ-NE for every k that satisfies

$$\sum_{\ell=1}^{k} \eta^\ell \geq O\Big(\frac{n \sum_{\ell=1}^{\infty} (\eta^\ell)^2 \sum_{i=1}^{n} |S_i||A_i|}{\alpha\epsilon}\Big).$$

Corollary

If we take $\eta^\ell = \ell^{-\frac{1}{2}-\beta}$ for some $\beta \to 0$, with probability at least $1 - \alpha$, $\bar{\rho}^k$ is an ϵ-NE for any $k \geq O\big(\frac{n}{\alpha\epsilon} \sum_{i=1}^{n} |S_i||A_i|\big)^2$.

Can we go beyond Social Concavity?

Definition: An occupation profile ρ^* is called a stable NE for the virtual game if $\langle \rho^* - \rho, v(\rho) \rangle \geq 0, \forall \rho \in \mathcal{P}$, with equality if and only if $\rho = \rho^*$, where $v(\rho) = (v_i(\rho_{-i}), i \in [n])$ is the vector of players' payoff gradients.

Theorem 4

Assume that the set of NE of the virtual game is stable. If each player follows Algorithm 1, under the same assumptions as before, almost surely the sequence of occupancy measures ρ^ℓ generated by the players converges to the set of stable ϵ-NE policies.

Remark: One can use the above theorem to obtain high-probability polynomial convergence rates. However, the convergence rates in this case are weaker in some sense compared to the case of social concavity.

Conclusions:

➢ We studied a subclass of stochastic games in which players have their own independent chains while they are coupled through their payoff functions.

➢ By establishing an equivalence between stationary NE policies in such games and NE points in a virtual game, we developed a scalable learning algorithm if

- the underlying virtual game is socially concave
- the underlying virtual game admits a stable NE.

Future directions:

❑ In general, there are strong computational lower bounds for developing scalable learning algorithms in n-player stochastic games.

❑ Stochastic games provide natural frameworks for modeling competition under uncertainty. What other classes of stochastic games admit scalable learning algorithms?

❑ One approach is to rely on mean-field approximations to simplify the learning task by allowing the players to focus on learning the mean-field trajectory of actions/states.

Thank You!