Independent Policy Mirror Descent for Markov Potential Games: Scaling to Large Number of Players

Anas Barakat*

Joint work with Pragnya Alatur and Niao He

* currently research fellow at Singapore University of Technology and Design

December 17, 2024- CDC 2024

ETH zürich

This Talk: Multi-Agent Reinforcement Learning

(a) Autonomous Driving

(b) Automated warehouse robots

(c) Smart Grids

(d) Communication Networks

Mathematical Framework for MARL

- **Stochastic Games** [Shapley, 1953]
  - $\Gamma = \langle \mathcal{N}, \mathcal{S}, (\mathcal{A}_i)_{i\in\mathcal{N}}, P, (r_i)_{i\in\mathcal{N}}, \rho, \gamma \rangle$
    - Here $P(s'|s, \mathbf{a})$, $r_i(s, \mathbf{a})$, dependence on **a** joint action (**strategic interaction**)
  - **joint** policy $\pi \in \Pi = \prod_{i\in\mathcal{N}} \Delta(\mathcal{A}_i)^{\mathcal{S}}$
  - (Action)-value functions for each agent $i \in \mathcal{N}, \pi \in \Pi$,

$$V_i^\pi(s) := \mathbb{E}_\pi \left[\sum_{t=0}^{+\infty} \gamma^t r_i(s_t, a_t) \mid s_0 = s \right], \quad Q_i^\pi(s, a) := \mathbb{E}_\pi \left[\sum_{t=0}^{+\infty} \gamma^t r_i(s_t, a_t) | s_0 = s, a_0 = a \right].$$

ε-approximate Nash equilibrium (ε-NE)

$$\pi^* = (\pi_i^*, \pi_{-i}^*) \in \Pi \quad \text{s.t.} \quad \forall i \in \mathcal{N}, \pi_i' \in \Delta(\mathcal{A}_i)^{\mathcal{S}}, \quad \mathbb{E}_{s\sim\rho}[V_i^{\pi^*}(s)] - \mathbb{E}_{s\sim\rho}[V_i^{\pi_i', \pi_{-i}^*}(s)] \leq \varepsilon$$

Outline

Independent Learning

- Learning protocol (see e.g. [Ozdaglar et al., 2021]), a.k.a. uncoupled learning
  - agents can **only observe realized state and their own reward and action** (e.g. not policies of others)
- Motivation
  - **Scaling** ('curse of multi-agents')
  - Privacy protection
  - Communication cost

Example: Dynamic load balancing [Yao and Ding, 2022]

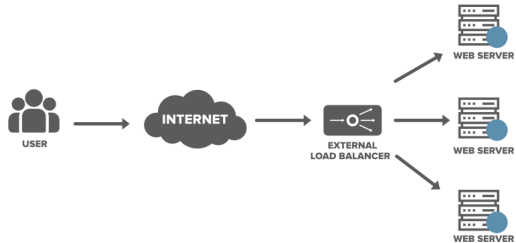

Figure 2: Source: geeksforgeeks.org

- Assign clients to servers in distributed computing
  - minimize communication overhead for low-latency response
  - scale across large data centers
- Can be modeled as a Markov Potential Game

Markov Potential Games

- Extend potential games in **static** normal form games (**approx NE tractable**).

Definition

$\forall s \in \mathcal{S}, \exists \Phi_s : \Pi \to \mathbb{R}$ (*player independent*) s.t. $\forall i \in \mathcal{N}$, $(\pi_i, \pi_{-i}) \in \Pi, \pi_i' \in \Delta(\mathcal{A}_i)^{\mathcal{S}}$,

$$V_i^{\pi_i, \pi_{-i}}(s) - V_i^{\pi_i', \pi_{-i}}(s) = \Phi_s(\pi_i, \pi_{-i}) - \Phi_s(\pi_i', \pi_{-i})$$

- Includes **identical interest case** ($r_i = r, \forall i$) and beyond.
- Actively investigated recently [Macua et al., 2018, Leonardos et al., 2022, Fox et al., 2022, Zhang et al., 2022b, Song et al., 2022, Ding et al., 2022, Zhang et al., 2022a, Maheshwari et al., 2023, Zhou et al., 2023].

Contributions in a nutshell

- **Policy Mirror Descent** algorithm implemented **independently** by agents:
  - Similarly to single agent setting.
  - Mirror descent with a dynamically weighted Bregman divergence regularization.
  - Unifies Policy Gradient Ascent and Natural Policy Gradient for MPGs in the lit.
- **Nash regret bounds** ($\implies$ ε-NE) for PMD with either Euclidean or KL reg.:
  - Setting: full information, i.e. access to average Q-values w.r.t. policies of others.
  - **From N to $\sqrt{N}$ dependence w.r.t. nb of players N using KL.**
  - Independence from the size of the agents' action spaces.
  - Improved state dependence in our MPG setting.

Independent PMD

For every state $s \in \mathcal{S}$, every agent $i \in \mathcal{N}$,

$$\pi_{i,s}^{(t+1)} \in \underset{\pi_{i,s} \in \Delta(\mathcal{A}_i)}{\operatorname{argmax}} \left\{ \langle \bar{Q}_{i,s}^{\pi^{(t)}}, \pi_{i,s} \rangle - \frac{1}{\eta} D_\psi(\pi_{i,s}, \pi_{i,s}^{(t)}) \right\}, \qquad \text{(PMD)}$$

- **averaged Q-value** $\bar{Q}_i^\pi : \mathcal{S} \times \mathcal{A}_i \to \mathbb{R}$ for any agent $i \in \mathcal{N}$ and policy $\pi \in \Pi$:

$$\bar{Q}_i^\pi(s, a_i) := \mathbb{E}_{a_{-i} \sim \pi_{-i}(\cdot|s)}[Q_i^\pi(s, a_i, a_{-i})], \ \forall s \in \mathcal{S}, \ a_i \in \mathcal{A}_i \, .$$

  - can be estimated independently by agents using only their observed rewards.

- **Bregman divergence** D_ψ induced by a mirror map $\psi : \operatorname{dom} \psi \to \mathbb{R}$ such that $\Delta(\mathcal{A}_i) \subset \operatorname{dom} \psi$, i.e. for any $p, q \in \Delta(\mathcal{A}_i)$,

$$D_\psi(p, q) = \psi(p) - \psi(q) - \langle \nabla \psi(q), p - q \rangle \, .$$

Examples of PMD for MPGs

Projected Q-Ascent

With mirror map $\psi(x) = \frac{1}{2}\|x\|^2$, $D_\psi(x, y) = \frac{1}{2}\|x - y\|^2$:

$$\pi_{i,s}^{(t+1)} = \mathrm{Proj}_{\Delta(\mathcal{A}_i)}(\pi_{i,s}^{(t)} + \eta \bar{Q}_{i,s}^{\pi^{(t)}}), \tag{1}$$

$\forall i \in \mathcal{N}, s \in \mathcal{S}$ and $\mathrm{Proj}_{\Delta(\mathcal{A}_i)}$ is the projection operator on the simplex $\Delta(\mathcal{A}_i)$.

Natural Policy Gradient/Exponential Q-Ascent

With $\psi =$ neg. entropy, $D_\psi =$ KL:

$$\pi_i^{(t+1)}(a_i|s) = \frac{\pi_i^{(t)}(a_i|s)\exp(\eta \bar{Q}_i^{\pi^{(t)}}(s, a_i))}{Z_t^{i,s}}, \tag{2}$$

$\forall i \in \mathcal{N}, (s, a_i) \in \mathcal{S} \times \mathcal{A}_i$, where $Z_t^{i,s}$ prob. normalization and $\pi^{(0)} \in int(\Delta(\mathcal{A})^{|\mathcal{S}|})$.

Nash Regret Analysis: Key = Potential Function Improvement

Definition

$$\text{Nash-regret(T)} \triangleq \frac{1}{T}\sum_{t=1}^{T} \max_{i\in\mathcal{N}} \max_{\pi_i'\in\Pi^i} V_i^{\pi_i',\pi_{-i}^{(t)}}(\rho) - V_i^{\pi^{(t)}}(\rho)\,,$$

Algorithm	ε-**Nash regret**
PMD (Euclidean reg.)	$\mathcal{O}\left(\frac{\phi_{\max}^2 \tilde{\kappa}_\rho \sum_{i=1}^{N} \lvert\mathcal{A}_i\rvert}{(1-\gamma)^4\varepsilon^2}\right)$
PMD (KL reg.)	$\mathcal{O}\left(\frac{\phi_{\max}^2 \tilde{\kappa}_\rho \sqrt{N}}{(1-\gamma)^4 c\varepsilon^2}\right)$

Table 1: Iteration complexity for obtaining ε-Nash regret in MPGs (c constant dependent on initial policy, $\tilde{\kappa}_\rho$ distrib. mismatch coeff., $\phi_{\max}$ max individual state potential fun.)

Future Work

- Technical follow-up on our work:
  - Stochastic setting (estimate average Q-functions)? **Stochastic approximation**!
  - Last iterate guarantees beyond average Nash regret?
  - Scaling to large spaces via function approximation?
- **More broadly, several interesting questions to investigate:**
  - **Centralized vs independent learning**: Can we characterize fundamental limits between these settings (e.g. in terms of iteration and sample complexity) for different algorithms/game dynamics?
  - **"Fully" independent learning** dynamics (agents running different algorithms)?
  - **Better definitions for cooperative MARL?** Definition of MPGs adopted in the lit. seems quite restrictive for fully capturing the dynamical stateful setting (as an extension of potential games). Can we go beyond?
  - **Other Classes of Markov Games with tractable NE computation? Equilibrium selection?** Active and exciting research areas.

Thank you for your attention

Check out our paper for more details:

Please feel free to reach out if you have any questions or comments!

References I

Ding, D., Wei, C.-Y., Zhang, K., and Jovanovic, M. (2022).
Independent Policy Gradient for Large-Scale Markov Potential Games: Sharper Rates, Function Approximation, and Game-Agnostic Convergence.
In *Proceedings of the 39th International Conference on Machine Learning*, pages 5166–5220. PMLR.
ISSN: 2640-3498.

Fox, R., Mcaleer, S. M., Overman, W., and Panageas, I. (2022).
Independent natural policy gradient always converges in markov potential games.
In *International Conference on Artificial Intelligence and Statistics*, pages 4414–4425. PMLR.

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. (2022).
Global convergence of multi-agent policy gradient in markov potential games.
In *International Conference on Learning Representations*.

References II

Macua, S. V., Zazo, J., and Zazo, S. (2018).
Learning parametric closed-loop policies for markov potential games.
In *International Conference on Learning Representations*.

Maheshwari, C., Wu, M., Pai, D., and Sastry, S. (2023).
Independent and Decentralized Learning in Markov Potential Games.
arXiv:2205.14590 [cs, eess].

Ozdaglar, A., Sayin, M. O., and Zhang, K. (2021).
Independent learning in stochastic games.
Invited chapter for the International Congress of Mathematicians 2022 (ICM 2022), arXiv preprint arXiv:2111.11743.

Shapley, L. S. (1953).
Stochastic games.
Proceedings of the national academy of sciences, 39(10):1095–1100.

References III

Song, Z., Mei, S., and Bai, Y. (2022).
When can we learn general-sum markov games with a large number of players sample-efficiently?
In *International Conference on Learning Representations*.

Yao, Z. and Ding, Z. (2022).
Learning distributed and fair policies for network load balancing as markov potential game.
Advances in Neural Information Processing Systems, 35:28815–28828.

Zhang, R., Mei, J., Dai, B., Schuurmans, D., and Li, N. (2022a).
On the global convergence rates of decentralized softmax gradient play in markov potential games.
Advances in Neural Information Processing Systems, 35:1923–1935.

References IV


Zhang, R. C., Ren, Z., and Li, N. (2022b).
Gradient play in stochastic games: Stationary points and local geometry.
IFAC-PapersOnLine, 55(30):73–78.
25th International Symposium on Mathematical Theory of Networks and Systems MTNS 2022.


Zhou, Z., Chen, Z., Lin, Y., and Wierman, A. (2023).
Convergence rates for localized actor-critic in networked Markov potential games.
In Evans, R. J. and Shpitser, I., editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2563–2573. PMLR.