Independent policy gradient for large-scale Markov potential games: sharper rates, function approximation, and game-agnostic convergence

Dongsheng Ding (USC)

a joint work with

Chen-Yu Wei (USC), Kaiqing Zhang (MIT), Mihailo R. Jovanović (USC)

39th International Conference on Machine Learning, Baltimore

Context

- Can many independent agents learn good policies?
  - ⋆ self-interested behaviors
  - ⋆ absence of first-principle models
  - ⋆ abundance of data

Success stories

StarCraft

Kilobots

Success stories

StarCraft

Kilobots

- Independent learning
  - ⋆ too many agents
  - ⋆ large state space

Success stories

StarCraft

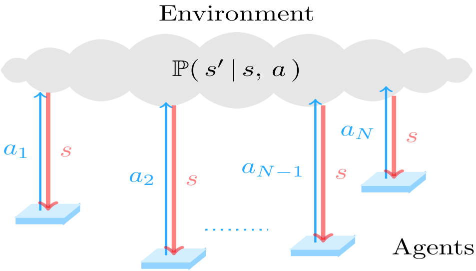
Kilobots

- Independent learning
  - ⋆ too many agents
  - ⋆ large state space

Question: provable scalability?

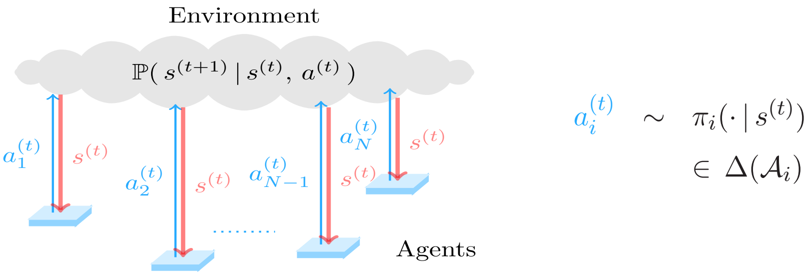
Markov game framework

Markov game

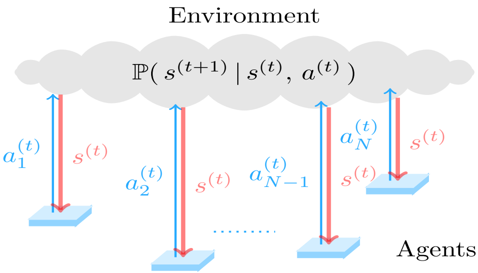


$$\big(\, \mathcal{S},\ \{\mathcal{A}_i\}_{i=1}^N,\ \mathbb{P},\ \{r_i\}_{i=1}^N,\ \rho \,\big)$$

- ⋆ $\mathcal{S}$ – state space; $\mathcal{A} := \prod_i \mathcal{A}_i$ – action space ($A = |\mathcal{A}_i|, \forall i$)
- ⋆ $\mathbb{P}(s' \mid s, a)$ – transition probability from s to s' given a
- ⋆ $r_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$ – agent i's reward function
- ⋆ ρ – initial state distribution

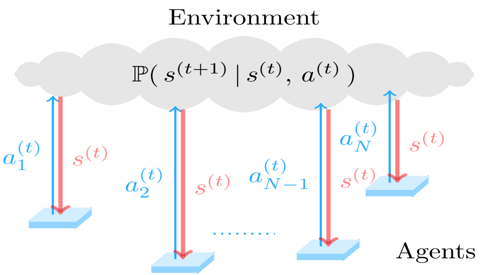


Environment
$\mathbb{P}(\, s^{(t+1)} \mid s^{(t)},\, a^{(t)} \,)$
$a_1^{(t)}$
$s^{(t)}$
$a_2^{(t)}$
$s^{(t)}$
$a_{N-1}^{(t)}$
$s^{(t)}$
$a_N^{(t)}$
$s^{(t)}$
Agents
$a_i^{(t)} \sim \pi_i(\cdot \mid s^{(t)})$
$\in \Delta(\mathcal{A}_i)$



$$a_i^{(t)} \sim \pi_i(\cdot \mid s^{(t)}) \in \Delta(\mathcal{A}_i)$$

$\star$ $V_i^{\pi}(s) := \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s^{(t)}, a^{(t)}) \mid s^{(0)} = s \right]$ – γ-discounted value



$$a_i^{(t)} \sim \pi_i(\cdot \mid s^{(t)}) \in \Delta(\mathcal{A}_i)$$

$\star$ $V_i^{\pi}(s) := \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s^{(t)}, a^{(t)}) \mid s^{(0)} = s \right]$ – γ-discounted value

- Nash policy $\pi^\star$

$$V_i^{\pi_i^\star, \pi_{-i}^\star}(s) \;\geq\; V_i^{\pi_i, \pi_{-i}^\star}(s), \quad \text{for all } s, \pi_i, \text{ and } i$$

usually exist, but hard to compute!

Daskalakis, TALG, ’13

Markov potential game

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

for any π_i, π_i', π_{-i}, and all i and s

$\Phi^{\pi}(s) : \Delta(\mathcal{A}) \times \mathcal{S} \to \mathbb{R}$ – (global) potential function

Markov potential game

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

for any π_i, π'_i, π_{-i}, and all i and s

$\Phi^{\pi}(s) : \Delta(\mathcal{A}) \times \mathcal{S} \to \mathbb{R}$ – (global) potential function

- Special cases
  - ⋆ common reward – Markov cooperative game
  - ⋆ stateless case – potential game (or congestion game)

Independent policy gradient ascent

$$
\begin{aligned}
\pi_1^{(t+1)} &\leftarrow \mathcal{P}_{\Delta(\mathcal{A}_1)^S}\left(\pi_1^{(t)} + \eta\,\nabla_{\pi_1} V_1^{(t)}\right)\\
\pi_2^{(t+1)} &\leftarrow \mathcal{P}_{\Delta(\mathcal{A}_2)^S}\left(\pi_2^{(t)} + \eta\,\nabla_{\pi_2} V_2^{(t)}\right)\\
&\vdots\\
\pi_N^{(t+1)} &\leftarrow \mathcal{P}_{\Delta(\mathcal{A}_N)^S}\left(\pi_N^{(t)} + \eta\,\nabla_{\pi_N} V_N^{(t)}\right)
\end{aligned}
$$

$S = |\mathcal{S}| < \infty$ – state space size

Leonardos, Overman, Panageas, Piliouras, ICLR, '22

Zhang, Ren, Li, arXiv:2106.00198, '21

Independent policy gradient ascent

$$\pi_1^{(t+1)} \leftarrow \mathcal{P}_{\Delta(\mathcal{A}_1)^S}\left(\pi_1^{(t)} + \eta \nabla_{\pi_1} V_1^{(t)}\right)$$

$$\pi_2^{(t+1)} \leftarrow \mathcal{P}_{\Delta(\mathcal{A}_2)^S}\left(\pi_2^{(t)} + \eta \nabla_{\pi_2} V_2^{(t)}\right)$$

$$\vdots$$

$$\pi_N^{(t+1)} \leftarrow \mathcal{P}_{\Delta(\mathcal{A}_N)^S}\left(\pi_N^{(t)} + \eta \nabla_{\pi_N} V_N^{(t)}\right)$$

$S = |\mathcal{S}| < \infty$ – state space size

Leonardos, Overman, Panageas, Piliouras, ICLR, '22

Zhang, Ren, Li, arXiv:2106.00198, '21

Restriction: small state space

Exponentially large state space

$$\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \ldots \times \mathcal{S}_N$$

$\mathcal{S}_i$ – agent i's state space

N – number of agents

e.g., Qu, Wierman, Li, L4DC, '20

Exponentially large state space

$$\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \ldots \times \mathcal{S}_N$$

$\mathcal{S}_i$ – agent i's state space

N – number of agents

e.g., Qu, Wierman, Li, L4DC, '20

- Scalability issue

$$|\mathcal{S}| = 2^{D \times N}$$

D – dimension of each agent's state

Can we design independent policy gradient methods for large-scale Markov games, with non-asymptotic convergence to a Nash policy?

Glimpse of our results (MPG)

$\boxed{\textbf{Performance:} \text{ an } \epsilon \text{ near-Nash policy}}$

Glimpse of our results (MPG)

Key feature: no explicit S-dependence

Methods	Iterations	Samples
Our method ★	$\frac{A\,N\,d^4}{\epsilon^2}$	$\frac{A^2\,N^2\,d^6}{\epsilon^5}$
Projected PG ascent ①	$\frac{S\,A\,N\,d^2}{\epsilon^2}$	$\frac{S^2\,A\,N\,d^4}{\epsilon^6}$
Projected PG ascent ②	$\frac{S\,A\,N\,\hat{d}^2}{\epsilon^2}$	$\frac{S^4\,A^3\,N\,\hat{d}^6}{\epsilon^6}$
Softmax PG ascent ③	$\frac{A\,N\,\tilde{d}^2}{c^2\,\epsilon^2}$	…

① Leonardos, et al, ICLR, '22

$d \;:=\; \sup_\pi \left\| d_\rho^\pi / \rho \right\|_\infty$

② Zhang, et al, arXiv, '21

$\hat{d} \;:=\; \sup_{\pi',\pi} \left\| d_\rho^{\pi'} / d_\rho^\pi \right\|_\infty$

③ Zhang, et al, arXiv, '22

$\tilde{d} \;:=\; \sup_\pi \left\| 1/d_\rho^\pi \right\|_\infty (\geq S)$

$c \;:=\; \min_{s,i,t} \pi_i^{(t)}(a_i^\star \mid s)$

Independent policy gradient ascent

(exact gradient)

Two pillars

- Q-value function

$$Q_i^\pi(s,a) \quad := \quad \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s^{(t)}, a^{(t)}) \,\middle|\, s^{(0)} = s, a^{(0)} = a\right]$$

Two pillars

- Q-value function

$$Q_i^{\pi}(s,a) \quad := \quad \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s^{(t)}, a^{(t)}) \,\middle|\, s^{(0)} = s, a^{(0)} = a\right]$$

  - $\star$ $\bar{Q}_i^{\pi}(s,a_i) := \mathbb{E}_{a_{-i}}\left[Q_i^{\pi_i,\pi_{-i}}(s,a_i,a_{-i})\right]$ – averaged Q

Two pillars

- Q-value function

$$Q_i^{\pi}(s,a) \;\; := \;\; \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s^{(t)}, a^{(t)}) \,\middle|\, s^{(0)} = s, a^{(0)} = a\right]$$

  ⋆ $\bar{Q}_i^{\pi}(s,a_i) := \mathbb{E}_{a_{-i}}\left[Q_i^{\pi_i,\pi_{-i}}(s,a_i,a_{-i})\right]$ – averaged Q

- State visitation distribution

$$d_{s^{(0)}}^{\pi}(s) \; = \; (1-\gamma)\sum_{t=0}^{\infty} \gamma^t P^{\pi}(s^{(t)} = s \,|\, s^{(0)})$$

Two pillars

- Q-value function

$$Q_i^{\pi}(s,a) \;:=\; \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r_i(s^{(t)},a^{(t)}) \,\middle|\, s^{(0)}=s, a^{(0)}=a\right]$$

  - ⋆ $\bar{Q}_i^{\pi}(s,a_i) := \mathbb{E}_{a_{-i}}\left[Q_i^{\pi_i,\pi_{-i}}(s,a_i,a_{-i})\right]$ – averaged Q

- State visitation distribution

$$d_{s^{(0)}}^{\pi}(s) \;=\; (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P^{\pi}(s^{(t)}=s \mid s^{(0)})$$

  - ⋆ $d_{\rho}^{\pi}(s) = \mathbb{E}_{s^{(0)}\sim\rho}\left[d_{s^{(0)}}^{\pi}(s)\right]$ – expectation over $s^{(0)}\sim\rho$

Vanilla policy gradient ascent

$$\pi_i^{(t+1)}(\cdot \mid s) \quad \leftarrow \quad \underset{\pi_i(\cdot \mid s) \in \Delta(\mathcal{A}_i)}{\operatorname{argmax}} \left(\langle \pi_i(\cdot \mid s), \nabla_{\pi_i} V_i^{(t)}(\rho) \rangle \; - \; \frac{1}{\eta} \mathcal{R}_s^{(t)} \right)$$

$\mathcal{R}_s^{(t)} = \frac{1}{2} \|\pi_i(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s)\|^2$ – L_2 regularization

$\nabla_{\pi_i(a_i \mid s)} V_i^{(t)}(\rho) = \frac{1}{1-\gamma} d_\rho^{(t)}(s) \bar{Q}_i^{(t)}(s, a_i)$ – policy gradient

Leonardos, Overman, Panageas, Piliouras, ICLR, '22

Zhang, Ren, Li, arXiv:2106.00198, '21

$$\pi_i^{(t+1)}(\cdot \mid s) \;\leftarrow\; \underset{\pi_i(\cdot \mid s) \in \Delta(\mathcal{A}_i)}{\operatorname{argmax}} \left(\langle \pi_i(\cdot \mid s), \nabla_{\pi_i} V_i^{(t)}(\rho) \rangle - \frac{1}{\eta} \mathcal{R}_s^{(t)} \right)$$

$\mathcal{R}_s^{(t)} = \frac{1}{2} d_\rho^{(t)}(s) \|\pi_i(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s)\|^2$ – weighted L_2 regularization

$\nabla_{\pi_i(a_i \mid s)} V_i^{(t)}(\rho) = \frac{1}{1-\gamma} d_\rho^{(t)}(s) \bar{Q}_i^{(t)}(s, a_i)$ – policy gradient

$$\pi_i^{(t+1)}(\cdot \mid s) \quad \leftarrow \quad \underset{\pi_i(\cdot \mid s) \in \Delta(\mathcal{A}_i)}{\operatorname{argmax}} \left(\langle \pi_i(\cdot \mid s), \nabla_{\pi_i} V_i^{(t)}(\rho) \rangle - \frac{1}{\eta} \mathcal{R}_s^{(t)} \right)$$

$\mathcal{R}_s^{(t)} = \frac{1}{2} \cancel{d_\rho^{(t)}(s)} \| \pi_i(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s) \|^2$ – weighted L_2 regularization

$\nabla_{\pi_i(a_i \mid s)} V_i^{(t)}(\rho) = \frac{1}{1-\gamma} \cancel{d_\rho^{(t)}(s)} \bar{Q}_i^{(t)}(s, a_i)$ – policy gradient

Independent Q-ascent

$$\pi_i^{(t+1)}(\cdot \mid s) \quad \leftarrow \quad \underset{\pi_i(\cdot \mid s) \in \Delta(\mathcal{A}_i)}{\operatorname{argmax}} \left(\langle \pi_i(\cdot \mid s), \nabla_{\pi_i} V_i^{(t)}(\rho) \rangle \; - \; \frac{1}{\eta} \mathcal{R}_s^{(t)} \right)$$

$\mathcal{R}_s^{(t)} = \frac{1}{2} \cancel{d_\rho^{(t)}(s)} \|\pi_i(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s)\|^2$ – weighted L_2 regularization

$\nabla_{\pi_i(a_i \mid s)} V_i^{(t)}(\rho) = \frac{1}{1-\gamma} \cancel{d_\rho^{(t)}(s)} \bar{Q}_i^{(t)}(s, a_i)$ – policy gradient

$$\Downarrow$$

$$\boxed{\pi_i^{(t+1)}(\cdot \mid s) \quad \leftarrow \quad \mathcal{P}_{\Delta(\mathcal{A}_i)}\left(\pi_i^{(t)}(\cdot \mid s) \; + \; \alpha \, \bar{Q}_i^{(t)}(s, \cdot) \right) \text{ for all } s, i}$$

Performance measure

- Nash regret

$$\text{Nash-Regret}(T) := \frac{1}{T}\sum_{t=1}^{T} \underbrace{\max_i \left(\max_{\pi_i'} V_i^{\pi_i', \pi_{-i}^{(t)}}(\rho) - V_i^{\pi^{(t)}}(\rho) \right)}_{\text{Nash gap}}$$

Performance measure

- Nash regret

$$\text{Nash-Regret}\,(T) \;:=\; \frac{1}{T}\sum_{t=1}^{T} \underbrace{\max_i \left(\max_{\pi_i'} V_i^{\pi_i', \pi_{-i}^{(t)}}(\rho) \;-\; V_i^{\pi^{(t)}}(\rho) \right)}_{\text{Nash gap}}$$

Objective: sublinear Nash-Regret (T), e.g., $\dfrac{1}{\sqrt{T}}$

Nash-Regret bound

Theorem (informal)

★ Markov potential game

$$\text{Nash-Regret}\,(T) \;\simeq\; d_p^2 \sqrt{\frac{A\,N}{T}}$$

★ Markov cooperative game

$$\text{Nash-Regret}\,(T) \;\simeq\; \sqrt{d_c}\sqrt{\frac{A\,N}{T}}$$

$d_p := \min(d, S)$ $\qquad$ $d_c := \min_\rho \left(d := \sup_\pi \, \|d_\rho^\pi/\rho\|_\infty \right)$

Nash-Regret bound

Theorem (informal)

★ Markov potential game

$$\text{Nash-Regret}\,(T) \;\simeq\; d_p^2 \sqrt{\frac{A\,N}{T}}$$

★ Markov cooperative game

$$\text{Nash-Regret}\,(T) \;\simeq\; \sqrt{d_c}\sqrt{\frac{A\,N}{T}}$$

$$d_p := \min(d, S) \qquad d_c := \min_\rho \left(d := \sup_\pi \; \|d_\rho^\pi/\rho\|_\infty\right)$$

- ⋆ sublinear regret & no explicit S-dependence
- ⋆ $d_c \le d_p \le d$ & d_c, $d_p < \infty$ for well-explored ρ

Independent policy gradient ascent

(no exact gradient, function approximation case)

Sample-based independent Q-ascent

- Linear averaged Q

$$\bar{Q}_i^\pi(s, a_i) \;:=\; \langle \phi_i(s, a_i), w_i^\pi \rangle, \quad \text{for all } \pi \text{ and } i$$

$\phi_i(s, a_i)$ – ith feature map with $\|\phi_i(s, a_i)\| \le 1$

$\|w_i^\pi\| \le W$ – bounded domain

Sample-based independent Q-ascent

- Linear averaged Q

$$\bar{Q}_i^\pi(s, a_i) \;:=\; \langle \phi_i(s, a_i), w_i^\pi \rangle, \quad \text{for all } \pi \text{ and } i$$

$\phi_i(s, a_i)$ – ith feature map with $\|\phi_i(s, a_i)\| \leq 1$

$\|w_i^\pi\| \leq W$ – bounded domain

$$\Downarrow$$

$$\boxed{\pi_i^{(t+1)}(\cdot \mid s) \;\leftarrow\; \mathcal{P}_{\Delta_\xi(\mathcal{A}_i)}\big(\pi_i^{(t)}(\cdot \mid s) + \alpha\, \hat{Q}_i^{(t)}(s, \cdot)\big) \;\text{ for all } s, i}$$

$\hat{Q}_i^{(t)}(s, \cdot)$ – local averaged Q-estimate

Agnostic Nash-Regret bound

Theorem (informal)

- ★ Markov potential game

$$\mathbb{E}\,[\,\text{Nash-Regret}\,(T)\,] \;\simeq\; d^2\sqrt{\frac{A\,N}{T}} + \sqrt[3]{d^2\,W\,A N\epsilon_{\text{stat}}}$$

- ★ Markov cooperative game

$$\mathbb{E}\,[\,\text{Nash-Regret}\,(T)\,] \;\simeq\; \sqrt{d}\sqrt{\frac{A\,N}{T}} + \sqrt[3]{d^2\,W\,A N\epsilon_{\text{stat}}}$$

$d := \sup_{\pi}\, \|d_{\rho}^{\pi}/\rho\|_{\infty}$

ϵ_{stat} – estimation error

Agnostic Nash-Regret bound

Theorem (informal)

- Markov potential game

$$\mathbb{E}\left[\,\text{Nash-Regret}\,(T)\,\right] \;\simeq\; d^2\sqrt{\frac{A\,N}{T}} + \sqrt[3]{d^2\,W\,AN\epsilon_{\text{stat}}}$$

- Markov cooperative game

$$\mathbb{E}\left[\,\text{Nash-Regret}\,(T)\,\right] \;\simeq\; \sqrt{d}\sqrt{\frac{A\,N}{T}} + \sqrt[3]{d^2\,W\,AN\epsilon_{\text{stat}}}$$

$d := \sup_\pi \; \|d_\rho^\pi/\rho\|_\infty$

ϵ_{stat} – estimation error

⋆ $\epsilon_{\text{stat}} \simeq \frac{1}{K}$ for K SGD steps leads to $TK \simeq \frac{1}{\epsilon^5}$ trajectory samples

Game-agnostic independent learning

(convergence in more than one type of games)

Independent optimistic Q-ascent

$$\boxed{\begin{aligned} \bar{\pi}_i^{(t+1)}(\cdot \,|\, s) \quad &\leftarrow \quad \mathcal{P}_{\Delta(\mathcal{A}_i)}\big(\,\bar{\pi}_i^{(t)}(\cdot \,|\, s) \,+\, \alpha\, \bar{\mathcal{Q}}_i^{(t)}(s,\cdot)\,\big) \\ \pi_i^{(t+1)}(\cdot \,|\, s) \quad &\leftarrow \quad \mathcal{P}_{\Delta(\mathcal{A}_i)}\big(\,\bar{\pi}_i^{(t+1)}(\cdot \,|\, s) \,+\, \alpha\, \bar{\mathcal{Q}}_i^{(t)}(s,\cdot)\,\big) \;\; \text{for all } s, i \end{aligned}}$$

$\bar{\mathcal{Q}}_i^{(t)}(s,\cdot)$ – smoothed critic

Wei, Lee, Zhang, Luo, COLT, '21

Independent optimistic Q-ascent

$$\begin{aligned} \bar{\pi}_i^{(t+1)}(\cdot\,|\,s) \;\; &\leftarrow \;\; \mathcal{P}_{\Delta(\mathcal{A}_i)}\big(\,\bar{\pi}_i^{(t)}(\cdot\,|\,s) \,+\, \alpha\,\bar{\mathcal{Q}}_i^{(t)}(s,\cdot)\,\big) \\ \pi_i^{(t+1)}(\cdot\,|\,s) \;\; &\leftarrow \;\; \mathcal{P}_{\Delta(\mathcal{A}_i)}\big(\,\bar{\pi}_i^{(t+1)}(\cdot\,|\,s) \,+\, \alpha\,\bar{\mathcal{Q}}_i^{(t)}(s,\cdot)\,\big) \;\; \text{for all } s, i \end{aligned}$$

$\bar{\mathcal{Q}}_i^{(t)}(s,\cdot)$ – smoothed critic

Wei, Lee, Zhang, Luo, COLT, '21

- Game-agnostic convergence

Theorem (informal)

★ Two-player Markov cooperative/competitive games

Last-iterate convergence & Nash-Regret $(T) \;\; \simeq^\star \;\; \dfrac{1}{T^{1/6}}$

Summary

- Independent policy gradient for MPG
  - ⋆ global convergence with no explicit S-dependence
  - ⋆ global convergence (up to an error) in function approximation case

- Independent optimistic policy gradient for Markov cooperative/competitive games
  - ⋆ game-agnostic convergence

Thank you for your attention.