Multi-Agent Reinforcement Learning for Large-Scale Markov Potential Games

Dongsheng Ding

Penn
UNIVERSITY of PENNSYLVANIA

joint work with:

Chen-Yu Wei

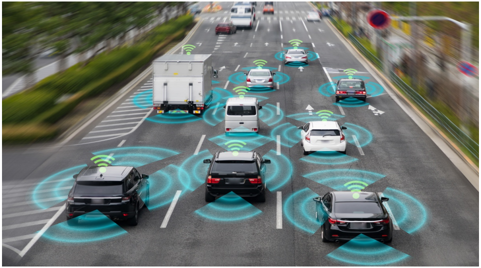
Kaiqing Zhang

Mihailo R. Jovanovic

TAMIDS Workshop on Multi-Agent Learning; April 26, 2024

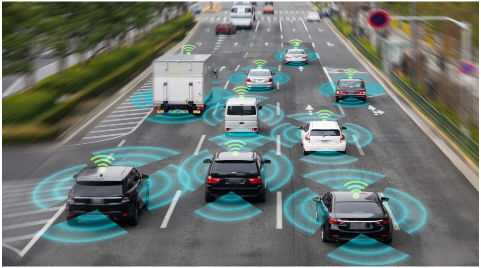
Motivating application

- MULTI-AGENT SYSTEM: CONNECTED VEHICLES



Motivating application

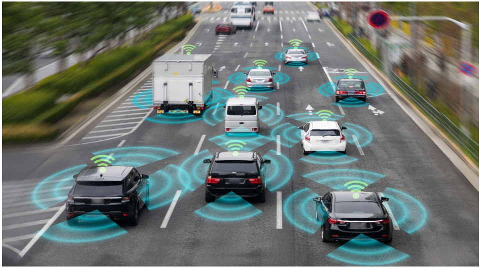
- MULTI-AGENT SYSTEM: CONNECTED VEHICLES



numerous agents make sequential decisions in unknown dynamic environment

Motivating application

- MULTI-AGENT SYSTEM: CONNECTED VEHICLES



numerous agents make sequential decisions in unknown dynamic environment

CHALLENGE: complexity of multiagency

Context

- SUCCESS STORIES OF RL

  Go/Atari game, drone/car racing, etc.

Context

- SUCCESS STORIES OF RL

  Go/Atari game, drone/car racing, etc.

- LESSONS LEARNED

  ⋆ importance of policy optimization

  simple; scalable; model-free

  ⋆ non-convex; optimality w/ one agent

  ⋆ difficult for many self-interest agents

  many solutions; non-stationary; stability; scalability

Context

- SUCCESS STORIES OF RL

  Go/Atari game, drone/car racing, etc.

- LESSONS LEARNED

  ⋆ importance of policy optimization

  simple; scalable; model-free

  ⋆ non-convex; optimality w/ one agent

  ⋆ difficult for many self-interest agents

  many solutions; non-stationary; stability; scalability

- WHAT NOW ?

  ⋆ application: multi-agent robotics, renewables, etc.

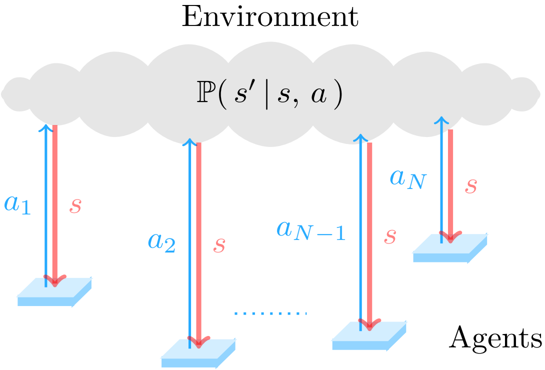
  ⋆ learning in games: tremendous advances

OBJECTIVE

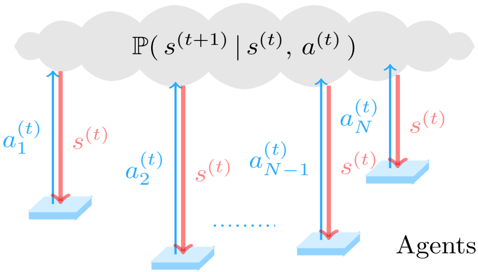
develop a learning strategy for multiple agents to interact w/ unknown dynamic environment, yielding a solution as an outcome, at scale

Framework of MARL

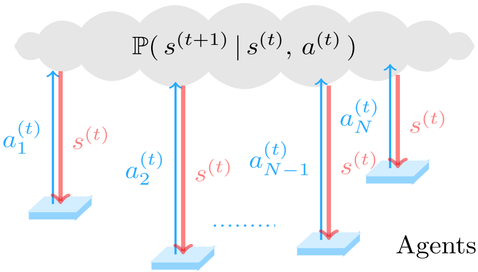
■ MARKOV GAME



- ⋆ ith policy $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$
- ⋆ transition probability $\mathbb{P}(s' \mid s, a)$
- ⋆ ith reward function $r_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$

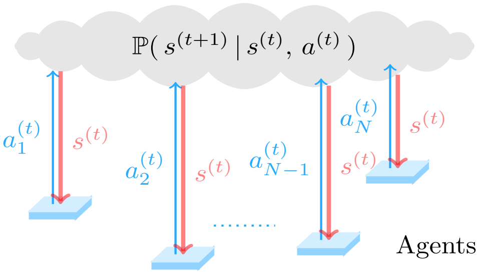


$$a_i^{(t)} \quad \sim \quad \pi_i(\cdot \mid s^{(t)})$$



$$a_i^{(t)} \sim \pi_i(\cdot \mid s^{(t)})$$

***i*th value function:** $$V_i^{\pi}(s) := \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s^{(t)}, a^{(t)}) \mid s^{(0)} = s\right]$$

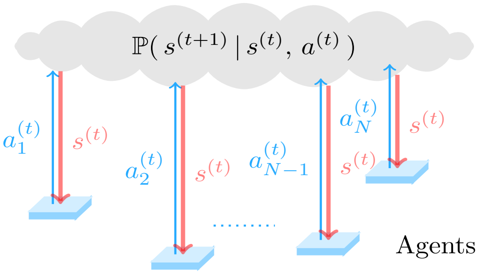


$$a_i^{(t)} \sim \pi_i(\cdot \mid s^{(t)})$$

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

NASH EQUILIBRIUM / POLICY

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



$$a_i^{(t)} \sim \pi_i(\cdot \mid s^{(t)})$$

***i*th value function:** $V_i^{\pi}(s) := \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s^{(t)}, a^{(t)}) \mid s^{(0)} = s\right]$

NASH EQUILIBRIUM / POLICY

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

often exists, but hard to compute

A glimpse of prior art

- TWO-PLAYER ZERO-SUM MARKOV GAME

Methods	Scalability	Stability
Centralized	☹	☺
Decentralized	☺	☹

A glimpse of prior art

- TWO-PLAYER ZERO-SUM MARKOV GAME

Methods	Scalability	Stability
Centralized	☹	☺
Decentralized	☺	☹

removal of coordination $\rightarrow$ nice scalability

non stationarity $\rightarrow$ poor stability

A glimpse of prior art

- TWO-PLAYER ZERO-SUM MARKOV GAME

Methods	Scalability	Stability
Centralized	☹	☺
Decentralized	☺	☹

removal of coordination → nice scalability

non stationarity → poor stability

CHALLENGE: scalability & stability

QUESTION

Can a Nash equilibrium of other Markov games be realized by decentralized methods ?

Empirical stories

StarCraft

Kilobots

all agents execute their own policy

and update rules w/o central controller

Empirical stories

StarCraft

Kilobots

all agents execute their own policy

and update rules w/o central controller

independent learning

Empirical stories

StarCraft

Kilobots

all agents execute their own policy

and update rules w/o central controller

independent learning

OBJECTIVE: provable guarantees

Markov potential game

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

Markov potential game

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

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

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

Markov potential game

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

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

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

- ⋆ Markov cooperative game
- ⋆ normal-form potential game

INDEPENDENT POLICY UPDATE

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

state space size $S = |\mathcal{S}|$

Leonardos, et al., ICLR, '22

Zhang, et al., arXiv:2106.00198, '23

INDEPENDENT POLICY UPDATE

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

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

$$\vdots$$

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

state space size $S = |\mathcal{S}|$

Leonardos, et al., ICLR, '22

Zhang, et al., arXiv:2106.00198, '23

find a near-Nash policy w/ explicit S-dependence

$\rightarrow$ unscalable in large state space

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

ith state space $\mathcal{S}_i$

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

ith state space $\mathcal{S}_i$

STATE SPACE SIZE

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

dimension of $\mathcal{S}_i$ = D

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

ith state space $\mathcal{S}_i$

STATE SPACE SIZE

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

dimension of $\mathcal{S}_i$ = D

exponentially large state space

CHALLENGE

independent learning methods for MPG

w/ numerous agents

& large state space

Overview of our results (MPG)

Performance: an ϵ near-Nash policy

Overview of our results (MPG)

Performance: an ϵ near-Nash policy

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

① Leonardos, et al, ICLR, '22 ② Zhang, et al, arXiv, '23 ③ Zhang, et al, NeurIPS, '22

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

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

Overview 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 gradient ①	$\frac{S\,A\,N\,d^2}{\epsilon^2}$	$\frac{S^2\,A\,N\,d^4}{\epsilon^6}$
Projected gradient ②	$\frac{S\,A\,N\,\hat{d}^2}{\epsilon^2}$	$\frac{S^4\,A^3\,N\,\hat{d}^6}{\epsilon^6}$
Softmax gradient ③	$\frac{A\,N\,\tilde{d}^2}{c^2\,\epsilon^2}$	...

① Leonardos, et al, ICLR, '22 ② Zhang, et al, arXiv, '23 ③ Zhang, et al, NeurIPS, '22

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

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

Independent policy gradient ascent

(exact gradient)

Two pillars

- Q-VALUE FUNCTION

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

Two pillars

- Q-VALUE FUNCTION

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

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

Two pillars

- Q-VALUE FUNCTION

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

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

- STATE VISITATION DISTRIBUTION

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

Two pillars

Q-VALUE FUNCTION

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

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

STATE VISITATION DISTRIBUTION

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

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

Vanilla policy gradient ascent

policy ascent direction: $\nabla_{\pi_i(a_i \,|\, s)} V_i^{(t)}(\rho) \;=\; \frac{1}{1-\gamma} d_\rho^{(t)}(s) \bar{Q}_i^{(t)}(s, a_i)$

L_2 **regularization:** $\mathcal{R}_s^{(t)} \;=\; \frac{1}{2} \|\pi_i(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s)\|^2$

Vanilla policy gradient ascent

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

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

- INDEPENDENT POLICY UPDATE

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

Vanilla policy gradient ascent

policy ascent direction: $\nabla_{\pi_i(a_i \,|\, s)} V_i^{(t)}(\rho) \;=\; \frac{1}{1-\gamma} d_\rho^{(t)}(s) \bar{Q}_i^{(t)}(s, a_i)$

L_2 **regularization:** $\mathcal{R}_s^{(t)} \;=\; \frac{1}{2} \|\pi_i(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s)\|^2$

- INDEPENDENT POLICY UPDATE

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

projected policy gradient ascent

Leonardos, et al., ICLR, ’22

Zhang, et al., arXiv:2106.00198, ’23

Independent Q-ascent

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

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

Independent Q-ascent

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

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

- INDEPENDENT POLICY UPDATE

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

Independent Q-ascent

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

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

- INDEPENDENT POLICY UPDATE

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

$$\boxed{\pi_i^{(t+1)}(\cdot \mid s) \;\leftarrow\; \mathcal{P}_{\Delta(\mathcal{A}_i)}\big(\pi_i^{(t)}(\cdot \mid s) + \eta \, \bar{Q}_i^{(t)}(s, \cdot) \big)}$$

projected Q-ascent

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, e.g., $\frac{1}{\sqrt{T}}$

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, e.g., $\frac{1}{\sqrt{T}}$

ϵ-Nash regret $\rightarrow$ **ϵ-Nash policy**

$$V_i^{\pi^{(t^\star)}}(\rho) \;\geq\; V_i^{\pi_i', \pi_{-i}^{(t^\star)}}(\rho) \;-\; \epsilon, \quad \text{for any } \pi_i' \text{ and } i$$

$$t^\star := \operatorname*{argmin}_{1 \leq t \leq T} \max_i \left(\max_{\pi_i'} V_i^{\pi_i', \pi_{-i}^{(t)}}(\rho) \;-\; V_i^{\pi^{(t)}}(\rho) \right)$$

Nash regret bound

Theorem (informal)

- ★ Markov potential game

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

- ★ Markov cooperative game

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

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

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 Nash regret

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 Nash regret

no explicit S-dependence

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 Nash regret

no explicit S-dependence

$\star$ $d_c \leq d_p \leq d$ & d_c, $d_p < \infty$ for well-explored ρ

Nash regret analysis (MPG)

Step #1: Performance difference & One-step optimality

$$\boxed{V_i^{\pi_i', \pi_{-i}^{(t)}}(\rho) - V_r^{\pi^{(t)}}(\rho)}$$

$$= \frac{1}{1-\gamma} \sum_{s,\, a_i} d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s) \Big(\pi_i'(a_i \,|\, s) - \pi_i^{(t)}(a_i \,|\, s) \Big) \bar{Q}_i^{(t)}(s, a_i)$$

$$\simeq \frac{1}{\eta} \sum_{s} d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\|$$

Nash regret analysis (MPG)

Step #1: Performance difference & One-step optimality

$$\boxed{V_i^{\pi_i', \pi_{-i}^{(t)}}(\rho) - V_r^{\pi^{(t)}}(\rho)}$$

$$= \frac{1}{1-\gamma} \sum_{s,\, a_i} d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s) \Big(\pi_i'(a_i \,|\, s) - \pi_i^{(t)}(a_i \,|\, s) \Big) \bar{Q}_i^{(t)}(s, a_i)$$

$$\simeq \frac{1}{\eta} \sum_{s} d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\|$$

Nash-Regret (T)

$$\boxed{\simeq \frac{1}{\eta\sqrt{T}} \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{s} d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\|^2}}$$

Step #2: Joint policy improvement

$\pi_1^{(t)}$

$\pi_2^{(t)}$

$\rightarrow$

$\pi_1^{(t+1)}$

$\pi_2^{(t+1)}$

$$\boxed{\Phi^{\pi^{(t+1)}} - \Phi^{\pi^{(t)}} = \mathsf{Diff}_1^{(t)} + \mathsf{Diff}_2^{(t)} + \mathsf{Cross}_{12}^{(t)}}$$

Step #2: Joint policy improvement

  $\rightarrow$  

$\pi_1^{(t)}$ $\pi_2^{(t)}$ $\pi_1^{(t+1)}$ $\pi_2^{(t+1)}$

$$\boxed{\Phi^{\pi^{(t+1)}} - \Phi^{\pi^{(t)}} = \text{Diff}_1^{(t)} + \text{Diff}_2^{(t)} + \text{Cross}_{12}^{(t)}}$$

$$\text{Diff}_i^{(t)} := \Phi^{\pi_i^{(t+1)},\pi_{-i}^{(t)}} - \Phi^{\pi^{(t)}} = V_i^{\pi_i^{(t+1)},\pi_{-i}^{(t)}} - V_i^{\pi^{(t)}}$$

$$\text{Cross}_{12}^{(t)} := \underbrace{\Phi^{\pi_1^{(t+1)},\pi_2^{(t+1)}} - \Phi^{\pi_1^{(t)},\pi_2^{(t+1)}} - \Phi^{\pi_1^{(t+1)},\pi_2^{(t)}} + \Phi^{\pi_1^{(t)},\pi_2^{(t)}}}_{V_1^{\pi_1^{(t+1)},\pi_2^{(t+1)}} - V_1^{\pi_1^{(t)},\pi_2^{(t+1)}} - V_1^{\pi_1^{(t+1)},\pi_2^{(t)}} + V_1^{\pi_1^{(t)},\pi_2^{(t)}}}$$

Step #2: Joint policy improvement

$\pi_1^{(t)}$

$\pi_2^{(t)}$

$\rightarrow$

$\pi_1^{(t+1)}$

$\pi_2^{(t+1)}$

$$\boxed{\Phi^{\pi^{(t+1)}} - \Phi^{\pi^{(t)}} = \text{Diff}_1^{(t)} + \text{Diff}_2^{(t)} + \text{Cross}_{12}^{(t)}}$$

$$\text{Diff}_i^{(t)} := \Phi^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}} - \Phi^{\pi^{(t)}} = V_i^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}} - V_i^{\pi^{(t)}}$$

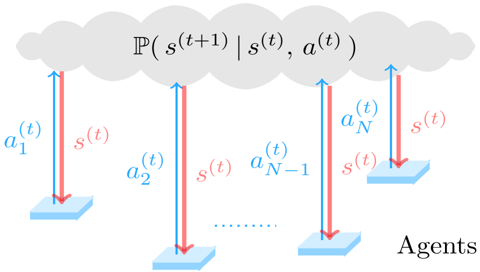
$$\text{Cross}_{12}^{(t)} := \underbrace{\Phi^{\pi_1^{(t+1)}, \pi_2^{(t+1)}} - \Phi^{\pi_1^{(t)}, \pi_2^{(t+1)}} - \Phi^{\pi_1^{(t+1)}, \pi_2^{(t)}} + \Phi^{\pi_1^{(t)}, \pi_2^{(t)}}}_{V_1^{\pi_1^{(t+1)}, \pi_2^{(t+1)}} - V_1^{\pi_1^{(t)}, \pi_2^{(t+1)}} - V_1^{\pi_1^{(t+1)}, \pi_2^{(t)}} + V_1^{\pi_1^{(t)}, \pi_2^{(t)}}}$$

$$\boxed{\eta \left(\Phi^{\pi^{(t+1)}} - \Phi^{\pi^{(t)}} \right) \simeq \sum_{i=1}^{N} \sum_{s} d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s) \right\|^2}$$

Independent policy gradient ascent

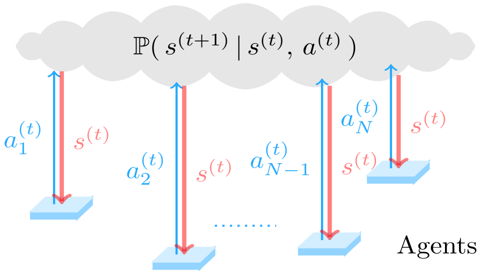
(no exact gradient, function approximation case)

Simulation setting



$$
\begin{aligned}
a_1^{(t)} &\sim \pi_1(\cdot \mid s^{(t)}) \\
a_2^{(t)} &\sim \pi_2(\cdot \mid s^{(t)}) \\
&\vdots \\
a_N^{(t)} &\sim \pi_N(\cdot \mid s^{(t)})
\end{aligned}
$$

Simulation setting

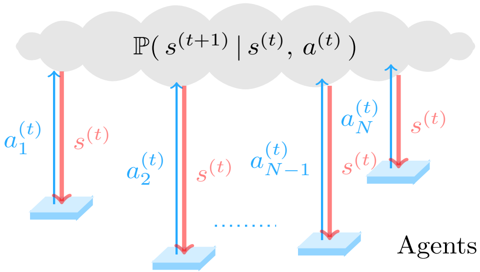


$$
\begin{aligned}
a_1^{(t)} &\sim \pi_1(\cdot \mid s^{(t)}) \\
a_2^{(t)} &\sim \pi_2(\cdot \mid s^{(t)}) \\
&\vdots \\
a_N^{(t)} &\sim \pi_N(\cdot \mid s^{(t)})
\end{aligned}
$$

- ⋆ trajectory of random horizon $\{(\bar{s}^{(h)}, \bar{a}^{(h)}, \bar{r}^{(h)})\}_{h=0}^{H-1}$

$$
\bar{a}^{(h)} = \left(\bar{a}_1^{(h)}, \bar{a}_2^{(h)}, \ldots, \bar{a}_N^{(h)}\right) \qquad \bar{r}^{(h)} = \left(\bar{r}_1^{(h)}, \bar{r}_2^{(h)}, \ldots, \bar{r}_N^{(h)}\right)
$$

Simulation setting



$$
\begin{aligned}
a_1^{(t)} &\sim \pi_1(\cdot \mid s^{(t)}) \\
a_2^{(t)} &\sim \pi_2(\cdot \mid s^{(t)}) \\
&\vdots \\
a_N^{(t)} &\sim \pi_N(\cdot \mid s^{(t)})
\end{aligned}
$$

- ⋆ trajectory of random horizon $\{(\bar{s}^{(h)}, \bar{a}^{(h)}, \bar{r}^{(h)})\}_{h=0}^{H-1}$

$$\bar{a}^{(h)} = \left(\bar{a}_1^{(h)}, \bar{a}_2^{(h)}, \ldots, \bar{a}_N^{(h)}\right) \qquad \bar{r}^{(h)} = \left(\bar{r}_1^{(h)}, \bar{r}_2^{(h)}, \ldots, \bar{r}_N^{(h)}\right)$$

- ⋆ unbiased estimate of $\bar{Q}_i^{(t)}(s, a_i)$: $R_i^{(k)} = \sum_{h=h_i}^{h_i+h_i'-1} \bar{r}_i^{(h)}$

Sample-based independent Q-ascent

averaged Q-estimate: $\hat{Q}_i^{(t)}$

ξ-exploration: $\Delta_\xi(\mathcal{A}_i) \;=\; \{(1-\xi)\pi_i(\cdot \mid s) + \xi \mathsf{Unif}_{\mathcal{A}_i}\}$

Sample-based independent Q-ascent

averaged Q-estimate: $\hat{Q}_i^{(t)}$

ξ-exploration: $\Delta_\xi(\mathcal{A}_i) \ = \ \{(1-\xi)\pi_i(\cdot\,|\,s) + \xi\mathsf{Unif}_{\mathcal{A}_i}\}$

- INDEPENDENT POLICY UPDATE

$$\pi_i^{(t+1)}(\cdot\,|\,s) \ \ \leftarrow \ \ \mathcal{P}_{\Delta_\xi(\mathcal{A}_i)}\big(\,\pi_i^{(t)}(\cdot\,|\,s) \,+\, \eta\,\hat{Q}_i^{(t)}(s,\cdot)\,\big)$$

Sample-based independent Q-ascent

averaged Q-estimate: $\hat{Q}_i^{(t)}$

ξ-exploration: $\Delta_\xi(\mathcal{A}_i) \;=\; \{(1-\xi)\pi_i(\cdot\,|\,s) + \xi\mathsf{Unif}_{\mathcal{A}_i}\}$

- INDEPENDENT POLICY UPDATE

$$\pi_i^{(t+1)}(\cdot\,|\,s) \;\;\leftarrow\;\; \mathcal{P}_{\Delta_\xi(\mathcal{A}_i)}\big(\,\pi_i^{(t)}(\cdot\,|\,s) \;+\; \eta\,\hat{Q}_i^{(t)}(s,\cdot)\,\big)$$

- LINEAR AVERAGED Q

$$\bar{Q}_i^\pi(s,a_i) \;=\; \langle \phi_i(s,a_i), w_i^\pi \rangle$$

ith feature $\phi_i(s,a_i)$

bounded domain $\|w_i^\pi\| \le W$

- ⋆ linear regression

$$\hat{w}_i^{(t)} \approx \underset{\|w_i\| \leq W}{\operatorname{argmin}} \sum_{k=1}^{K} \left(R_i^{(k)} - w_i^\top \phi_i(s_i^{(k)}, a_i^{(k)}) \right)^2$$

⋆ linear regression

$$\hat{w}_i^{(t)} \;\approx\; \underset{\|w_i\| \leq W}{\operatorname{argmin}} \sum_{k=1}^{K} \Big(R_i^{(k)} - w_i^\top \phi_i(s_i^{(k)}, a_i^{(k)}) \Big)^2$$

INDEPENDENT POLICY UPDATE

$$\pi_i^{(t+1)}(\cdot \,|\, s) \;\leftarrow\; \mathcal{P}_{\Delta_\xi(\mathcal{A}_i)}\big(\pi_i^{(t)}(\cdot \,|\, s) \,+\, \eta \, \langle \phi_i(s, \cdot), \hat{w}_i^{(t)} \rangle \big)$$

⋆ linear regression

$$\hat{w}_i^{(t)} \;\approx\; \underset{\|w_i\| \leq W}{\operatorname{argmin}} \sum_{k=1}^{K} \Big(R_i^{(k)} - w_i^\top \phi_i(s_i^{(k)}, a_i^{(k)}) \Big)^2$$

INDEPENDENT POLICY UPDATE

$$\pi_i^{(t+1)}(\cdot \,|\, s) \;\leftarrow\; \mathcal{P}_{\Delta_\xi(\mathcal{A}_i)}\big(\, \pi_i^{(t)}(\cdot \,|\, s) \,+\, \eta \, \langle \phi_i(s, \cdot), \hat{w}_i^{(t)} \rangle \,\big)$$

sample-based projected Q-ascent

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

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

estimation error ϵ_{stat}

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

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

estimation error ϵ_{stat}

sublinear Nash regret (up to an error)

no explicit S-dependence

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

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

estimation error ϵ_{stat}

sublinear Nash regret (up to an error)

no explicit S-dependence

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

Game-agnostic independent learning

(convergence in more than one type of games)

Independent optimistic Q-ascent

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

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

Wei, et al., COLT, '21

Independent optimistic Q-ascent

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

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

Wei, et al., COLT, '21

■ GAME-AGNOSTIC CONVERGENCE

Theorem (informal)

★ Two-player Markov cooperative/competitive games

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

Summary

- INDEPENDENT Q-ASCENT
  - ⋆ global convergence with no explicit S-dependence
  - ⋆ global convergence in function approximation case
- INDEPENDENT OPTIMISTIC Q-ASCENT
  - ⋆ game-agnostic convergence

Ding, et al., ICML, ’22

Future directions

- BEYOND MARKOV POTENTIAL GAMES
  - ⋆ other potential games
  - ⋆ game-agnostic convergence
- CONSTRAINED MULTI-AGENT SYSTEMS
  - ⋆ constrained Markov games

Thank you for your attention.