Independent Learning in Constrained Markov Potential Games

Anas Barakat

Joint work with Philip Jordan and Niao He

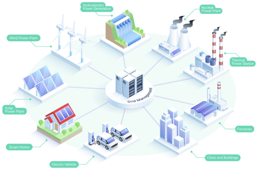
May 3rd

ETH zürich

Multi-Agent Reinforcement Learning

(a) Autonomous Driving

(b) Automated warehouse robots



(c) Smart Grids

(d) Communication Networks

Constraints in MARL

- Why constraints?
  - Physical system constraints
  - Safety considerations
  - ...
- Type of constraints?
  - 'Hard' constraints
    - e.g. collision avoidance
  - 'Soft' constraints: approximately satisfying the constraints can be tolerated
    - average user's total latency thresholds in wireless networks
    - average power constraints in signal transmission

Mathematical Framework

- Stochastic Games [Shapley, 1953]
  - $\mathcal{G} = (\mathcal{S}, \mathcal{N}, \{\mathcal{A}_i, r_i\}_{i \in \mathcal{N}}, \mu, P, \kappa)$
  - **joint** policy $\pi \in \Pi = \prod_{i \in \mathcal{N}} \Delta(\mathcal{A}_i)^{\mathcal{S}}$
  - Value function for each agent $i \in \mathcal{N}$

$$V_{r_i}(\pi) := \mathbb{E}_{s \sim \mu} \left[\sum_{t=0}^{T} r_i(s_t, a_t) \mid s_0 = s \right]$$

- Constrained Markov Games [Altman and Shwartz, 2000]
  - cost functions $c_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$ for each agent $i \in \mathcal{N}$
  - Thresholds α_i

Outline

1. Motivation and Problem Formulation
    - Independent Learning
    - MPGs
    - CMPGs
2. Related Work & Challenges
3. Algorithm
4. Iteration and Sample Complexity Analysis
5. Simulations: Distributed Energy Marketplace

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
- Motivation
  - Scaling ('curse of multi-agents')
  - Privacy protection
  - Communication cost

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

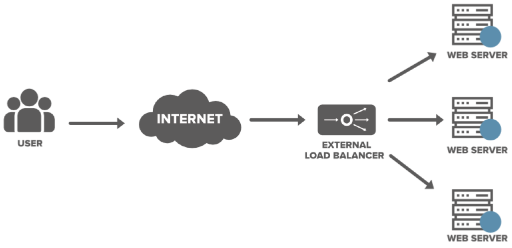


Figure 2: Source: geeksforgeeks.org

- Assign clients to servers in distributed computing
  - minimize communication overhead for low-latency response
  - scale across large data centers
- Can be modelled as an MPG

Markov Potential Games

- Extension of potential games

Definition

$\forall s \in \mathcal{S}, \exists \Phi_s : \Pi \to \mathbb{R}$ s.t. $\forall i \in \mathcal{N}$, $(\pi_i, \pi_{-i}) \in \Pi$, and $\pi'_i \in \Pi'_i$,

$$V_{r_i,s}(\pi_i, \pi_{-i}) - V_{r_i,s}(\pi'_i, \pi_{-i}) = \Phi_s(\pi_i, \pi_{-i}) - \Phi_s(\pi'_i, \pi_{-i})$$

- includes identical interest case 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].

ϵ-approximate Nash equilibrium (ϵ-NE)

$$\pi^* \in \Pi \quad \text{s.t.} \quad \forall i \in \mathcal{N}, \pi'_i \in \Pi^i, V_{r_i}(\pi^*) - V_{r_i}(\pi'_i, \pi^*_{-i}) \leq \epsilon.$$

Constrained Markov Potential Games

- subset of feasible policies $\Pi_c := \{\pi \in \Pi \mid V_c(\pi) \leq \alpha\}$; $\alpha \in \mathbb{R}$,

$$V_c(\pi) := \mathbb{E}_{s_0 \sim \mu}\left[\sum_{t=0}^{T} c(s_t, a_t)\right]$$

- **Here, same cost function for all agents, other case more challenging**

ϵ-approximate constrained NE

$$\pi^* \in \Pi_c \quad \text{s.t.} \quad \forall i \in \mathcal{N}, \pi_i' \in \Pi_c^i(\pi_{-i}^*), \quad V_{r_i}(\pi^*) - V_{r_i}(\pi_i', \pi_{-i}^*) \leq \epsilon.$$

Outline

Related Work

	centralized	**independent**
MPG	Nash-CA [Song et al., 2022]	Independent PGA [Leonardos et al., 2022],[Zhang et al., 2022b] [Ding et al., 2022]
CMPG	CA-CMPG [Alatur et al., 2023]	**?**

Related Work (centralized setting)

- **Nash-CA for MPGs [Song et al., 2022]**
  - Turn-based, fix $\pi_{-i}^{(t)}$
  - Solve an MDP computing a best response policy

$$\hat{\pi}_i^{(t+1)} = \arg\max_{\pi_i \in \Pi^i} V_{r_i}(\pi_i, \pi_{-i}^{(t)})$$

- **Nash-CA for CMPGs [Alatur et al., 2023]**
  - Turn-based, fix $\pi_{-i}^{(t)}$
  - Solve a CMDP computing a best response policy

$$\hat{\pi}_i^{(t+1)} = \arg\max_{\pi_i \in \Pi_c^i(\pi_{-i}^{(t)})} V_{r_i}(\pi_i, \pi_{-i}^{(t)})$$

Related Work (independent learning)

- **Independent PGA [Leonardos et al., 2022]**

  Simultaneously $\forall i \in \mathcal{N}$,

  $$\pi_i^{(t+1)} = \mathcal{P}_{\Pi^i}\left[\pi_i^{(t)} - \eta \nabla_{\pi_i} V_{r_i}(\pi^{(t)})\right]$$

  $$\pi^{(t+1)} = \mathcal{P}_{\Pi}\left[\pi^{(t)} - \eta \nabla_{\pi} \Phi(\pi^{(t)})\right]$$

  - ϵ-stationary point of Φ is $\mathcal{O}(\epsilon)$-NE

Challenges

- (no centralization) Environment is non-stationary from the viewpoint of each agent

$$\min_{(\pi_1,\ldots,\pi_m)\in\Pi_c} \Phi(\pi) \quad ; \quad \Pi_c := \{\pi \in \Pi \mid V_c(\pi) \le \alpha\}$$

- nonconvex objective *and* constraint
- constraint *couples* π_i's
- strong duality *does not* hold [Alatur et al., 2023]

Outline

1. Motivation and Problem Formulation
   - Independent Learning
   - MPGs
   - CMPGs
2. Related Work & Challenges
3. Algorithm
4. Iteration and Sample Complexity Analysis
5. Simulations: Distributed Energy Marketplace

Our Approach

$$\min_{(\pi_1,\ldots,\pi_m)\in\Pi_c} \Phi(\pi) \quad ; \Pi_c := \{\pi \in \Pi \mid V_c(\pi) \leq \alpha\} \tag{1}$$

Lemma

If π is an ϵ-KKT policy of (1), then π is a constrained $\mathcal{O}(\epsilon)$-NE.

- How to find an ϵ-KKT policy?

How to find ϵ-KKT policy?

- proximal-point-like update [Boob et al., 2023, Ma et al., 2020, Jia and Grimmer, 2023]

$$\pi^{(t+1)} = \arg\min_{\pi \in \Pi} \left\{ \Phi(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \;\middle|\; V_c(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \leq \alpha \right\}$$

How to find ϵ-KKT policy?

- proximal-point-like update
  [Boob et al., 2023, Ma et al., 2020, Jia and Grimmer, 2023]

$$\pi^{(t+1)} = \operatorname*{arg\,min}_{\pi \in \Pi} \left\{ \Phi(\pi) + \frac{1}{2\eta} \|\pi - \pi^{(t)}\|^2 \;\middle|\; V_c(\pi) + \frac{1}{2\eta} \|\pi - \pi^{(t)}\|^2 \leq \alpha \right\}$$

How to find ϵ-KKT policy?

- proximal-point-like update
  [Boob et al., 2023, Ma et al., 2020, Jia and Grimmer, 2023]

$$\pi^{(t+1)} = \operatorname*{arg\,min}_{\pi \in \Pi} \left\{ \Phi(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \;\middle|\; V_c(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \leq \alpha \right\}$$

- Φ and V_c weakly convex $\Rightarrow$ subproblem obj. and constr. strongly convex

How to find ϵ-KKT policy?

- proximal-point-like update [Boob et al., 2023, Ma et al., 2020, Jia and Grimmer, 2023]
$$\pi^{(t+1)} = \operatorname*{arg\,min}_{\pi \in \Pi} \left\{ \Phi(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \;\middle|\; V_c(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \le \alpha \right\}$$
- Φ and V_c weakly convex $\Rightarrow$ subproblem obj. and constr. strongly convex
- as $\left\|\pi^{(t+1)} - \pi^{(t)}\right\| \to 0$, regularized constraint approaches original constraint

How to find ϵ-KKT policy?

- proximal-point-like update
  [Boob et al., 2023, Ma et al., 2020, Jia and Grimmer, 2023]
  $$\pi^{(t+1)} = \arg\min_{\pi\in\Pi} \left\{ \Phi(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \;\middle|\; V_c(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \le \alpha \right\}$$
- Φ and V_c weakly convex $\Rightarrow$ subproblem obj. and constr. strongly convex
- as $\left\|\pi^{(t+1)} - \pi^{(t)}\right\| \to 0$, regularized constraint approaches original constraint

Can show:

$$\left\|\pi^{(t+1)} - \pi^{(t)}\right\| \le \epsilon \quad \Longrightarrow \quad \pi^{(t+1)} \text{ is } \mathcal{O}(\epsilon)\text{-KKT for } \min_{(\pi_1,\ldots,\pi_m)\in\Pi_c} \Phi(\pi)$$

How to find ϵ-KKT policy?

- proximal-point-like update
  [Boob et al., 2023, Ma et al., 2020, Jia and Grimmer, 2023]
  $$\pi^{(t+1)} = \arg\min_{\pi \in \Pi} \left\{ \Phi(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \;\middle|\; V_c(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \le \alpha \right\}$$
- Φ and V_c weakly convex $\Rightarrow$ subproblem obj. and constr. strongly convex
- as $\left\|\pi^{(t+1)} - \pi^{(t)}\right\| \to 0$, regularized constraint approaches original constraint

Can show:

$$\left\|\pi^{(t+1)} - \pi^{(t)}\right\| \le \epsilon \quad \Longrightarrow \quad \pi^{(t+1)} \text{ is } \mathcal{O}(\epsilon)\text{-KKT for } \min_{(\pi_1,\dots,\pi_m)\in\Pi_c} \Phi(\pi)$$

$$\overset{Lem.\ 1}{\Longrightarrow} \quad \pi^{(t+1)} \text{ is constrained } \mathcal{O}(\epsilon)\text{-NE}$$

How to find ϵ-KKT policy?

- proximal-point-like update
  [Boob et al., 2023, Ma et al., 2020, Jia and Grimmer, 2023]
  $$\pi^{(t+1)} = \arg\min_{\pi \in \Pi} \left\{ \Phi(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \;\middle|\; V_c(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \leq \alpha \right\}$$
- Φ and V_c weakly convex $\Rightarrow$ subproblem obj. and constr. strongly convex
- as $\left\|\pi^{(t+1)} - \pi^{(t)}\right\| \to 0$, regularized constraint approaches original constraint

Can show:

$$\left\|\pi^{(t+1)} - \pi^{(t)}\right\| \leq \epsilon \quad \Longrightarrow \quad \pi^{(t+1)} \text{ is } \mathcal{O}(\epsilon)\text{-KKT for } \min_{(\pi_1,\ldots,\pi_m)\in\Pi_c} \Phi(\pi)$$

$$\xRightarrow{\textit{Lem. } 1} \quad \pi^{(t+1)} \text{ is constrained } \mathcal{O}(\epsilon)\text{-NE}$$

- How to solve the proximal-point subproblem?

How to solve the proximal-point subproblem?

$$\pi^{(t+1)} = \operatorname*{arg\,min}_{\pi \in \Pi} \left\{ \Phi\left(\pi\right) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \;\middle|\; V_c(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2 \leq \alpha \right\}$$

How to solve the proximal-point subproblem?

$$\pi^{(t+1)} = \arg\min_{\pi \in \Pi} \left\{ \underbrace{\Phi(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2}_{=: \Phi_{\eta,\pi^{(t)}}(\pi)} \;\middle|\; \underbrace{V_c(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2}_{=: V^c_{\eta,\pi^{(t)}}(\pi)} \le \alpha \right\}$$

How to solve the proximal-point subproblem?

$$\pi^{(t+1)} = \arg\min_{\pi\in\Pi} \Big\{ \underbrace{\Phi(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2}_{=:\,\Phi_{\eta,\pi^{(t)}}(\pi)} \;\Big|\; \underbrace{V_c(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2}_{=:\,V^c_{\eta,\pi^{(t)}}(\pi)} \le \alpha \Big\}$$

- Solve via gradient switching subroutine [Lan and Zhou, 2020]:

$$\pi^{(t,k+1)} = \begin{cases} \mathcal{P}_\Pi\left[\pi^{(t,k)} - \nu_k \nabla_\pi \Phi_{\eta,\pi^{(t,k)}}\big(\pi^{(t,k)}\big)\right] & \text{if } V^c_{\eta,\pi^{(t,k)}}\big(\pi^{(t,k)}\big) - \alpha \le \delta_k, \\ \mathcal{P}_\Pi\left[\pi^{(t,k)} - \nu_k \nabla_\pi V^c_{\eta,\pi^{(t,k)}}\big(\pi^{(t,k)}\big)\right] & \text{otherwise} \end{cases}$$

How to solve the proximal-point subproblem?

$$\pi^{(t+1)} = \underset{\pi \in \Pi}{\arg\min} \left\{ \underbrace{\Phi(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2}_{=: \Phi_{\eta,\pi^{(t)}}(\pi)} \;\middle|\; \underbrace{V_c(\pi) + \frac{1}{2\eta}\|\pi - \pi^{(t)}\|^2}_{=: V^c_{\eta,\pi^{(t)}}(\pi)} \leq \alpha \right\}$$

- Solve via gradient switching subroutine [Lan and Zhou, 2020]:

$$\pi^{(t,k+1)} = \begin{cases} \mathcal{P}_\Pi \left[\pi^{(t,k)} - \nu_k \nabla_\pi \Phi_{\eta,\pi^{(t,k)}}\left(\pi^{(t,k)}\right) \right] & \text{if } V^c_{\eta,\pi^{(t,k)}}\left(\pi^{(t,k)}\right) - \alpha \leq \delta_k, \\ \mathcal{P}_\Pi \left[\pi^{(t,k)} - \nu_k \nabla_\pi V^c_{\eta,\pi^{(t,k)}}\left(\pi^{(t,k)}\right) \right] & \text{otherwise} \end{cases}$$

- independent implementation?

Independent implementation

- **Observation:**

$$\nabla_{\pi_i} \Phi_{\eta,\pi'}(\pi) = \nabla_{\pi_i} \Phi(\pi) + \frac{1}{\eta}\left(\pi_i - \pi_i'\right) = \nabla_{\pi_i} V_{r_i}(\pi) + \frac{1}{\eta}\left(\pi_i - \pi_i'\right)$$

Independent implementation

- **Observation:**

$$\nabla_{\pi_i}\Phi_{\eta,\pi'}(\pi) = \nabla_{\pi_i}\Phi(\pi) + \frac{1}{\eta}\left(\pi_i - \pi_i'\right) = \nabla_{\pi_i} V_{r_i}(\pi) + \frac{1}{\eta}\left(\pi_i - \pi_i'\right)$$

$\Rightarrow$ gradient switching update

$$\pi^{(t,k+1)} = \begin{cases} \mathcal{P}_\Pi\left[\pi^{(t,k)} - \nu_k \nabla_\pi \Phi_{\eta,\pi^{(t,k)}}\left(\pi^{(t,k)}\right)\right] & \text{if } V^c_{\eta,\pi^{(t,k)}}\left(\pi^{(t,k)}\right) - \alpha \le \delta_k, \\ \mathcal{P}_\Pi\left[\pi^{(t,k)} - \nu_k \nabla_\pi V^c_{\eta,\pi^{(t,k)}}\left(\pi^{(t,k)}\right)\right] & \text{otherwise} \end{cases}$$

Independent implementation

- **Observation:**

$$\nabla_{\pi_i}\Phi_{\eta,\pi'}(\pi) = \nabla_{\pi_i}\Phi(\pi) + \frac{1}{\eta}\left(\pi_i - \pi_i'\right) = \nabla_{\pi_i}V_{r_i}(\pi) + \frac{1}{\eta}\left(\pi_i - \pi_i'\right)$$

$\Rightarrow$ gradient switching update

$$\pi^{(t,k+1)} = \begin{cases} \mathcal{P}_{\Pi}\left[\pi^{(t,k)} - \nu_k \nabla_\pi \Phi_{\eta,\pi^{(t,k)}}\left(\pi^{(t,k)}\right)\right] & \text{if } V^c_{\eta,\pi^{(t,k)}}\left(\pi^{(t,k)}\right) - \alpha \le \delta_k, \\ \mathcal{P}_{\Pi}\left[\pi^{(t,k)} - \nu_k \nabla_\pi V^c_{\eta,\pi^{(t,k)}}\left(\pi^{(t,k)}\right)\right] & \text{otherwise} \end{cases}$$

is equivalent to independently, for all $i \in \mathcal{N}$,

$$\pi_i^{(t,k+1)} = \begin{cases} \mathcal{P}_{\Pi^i}\left[\pi_i^{(t,k)} - \nu_k \nabla_{\pi_i} V_{r_i}\left(\pi^{(t,k)}\right) - \frac{\nu_k}{\eta}\left(\pi_i^{(t,k)} - \pi_i^{(t)}\right)\right] \\ \qquad \text{if } V_c\left(\pi^{(t,k)}\right) + \frac{1}{2\eta}\left\|\pi^{(t,k)} - \pi^{(t)}\right\|^2 - \alpha \le \delta_k \\ \mathcal{P}_{\Pi^i}\left[\pi_i^{(t,k)} - \nu_k \nabla_{\pi_i} V_c\left(\pi^{(t,k)}\right) - \frac{\nu_k}{\eta}\left(\pi_i^{(t,k)} - \pi_i^{(t)}\right)\right] \\ \qquad \text{otherwise} \end{cases}$$

Independent implementation

- **Observation:**

$$\nabla_{\pi_i}\Phi_{\eta,\pi'}(\pi) = \nabla_{\pi_i}\Phi(\pi) + \frac{1}{\eta}\left(\pi_i - \pi_i'\right) = \nabla_{\pi_i} V_{r_i}(\pi) + \frac{1}{\eta}\left(\pi_i - \pi_i'\right)$$

$\Rightarrow$ gradient switching update

$$\pi^{(t,k+1)} = \begin{cases} \mathcal{P}_\Pi\left[\pi^{(t,k)} - \nu_k \nabla_\pi \Phi_{\eta,\pi^{(t,k)}}\left(\pi^{(t,k)}\right)\right] & \text{if } V^c_{\eta,\pi^{(t,k)}}\left(\pi^{(t,k)}\right) - \alpha \le \delta_k, \\ \mathcal{P}_\Pi\left[\pi^{(t,k)} - \nu_k \nabla_\pi V^c_{\eta,\pi^{(t,k)}}\left(\pi^{(t,k)}\right)\right] & \text{otherwise} \end{cases}$$

is equivalent to independently, for all $i \in \mathcal{N}$,

$$\pi_i^{(t,k+1)} = \begin{cases} \mathcal{P}_{\Pi^i}\left[\pi_i^{(t,k)} - \nu_k \nabla_{\pi_i} V_{r_i}\left(\pi^{(t,k)}\right) - \frac{\nu_k}{\eta}\left(\pi_i^{(t,k)} - \pi_i^{(t)}\right)\right] \\ \qquad \text{if } V_c\left(\pi^{(t,k)}\right) \cancel{+ \frac{1}{2\eta}\left\|\pi^{(t,k)} - \pi^{(t)}\right\|^2} - \alpha \le \delta_k \\ \mathcal{P}_{\Pi^i}\left[\pi_i^{(t,k)} - \nu_k \nabla_{\pi_i} V_c\left(\pi^{(t,k)}\right) - \frac{\nu_k}{\eta}\left(\pi_i^{(t,k)} - \pi_i^{(t)}\right)\right] \\ \qquad \text{otherwise} \end{cases}$$

Algorithm 1 iProxCMPG: independent **Prox**imal-policy algorithm for **CMPG**s

initialization: $\pi^{(0)} \in \Pi^{\xi}$ s.t. $V_c(\pi^{(0)}) < \alpha$ and suitably chosen $\eta, \xi, T, K, \{(\nu_k, \delta_k)\}_{0 \le k \le K}$

for $t = 0, \ldots, T-1$ **do**

  $\pi_i^{(t,0)} = \pi_i^{(t)}$

proximal update

  for $k = 0, \ldots, K-1$ and $i \in \mathcal{N}$ simultaneously **do**

    sample B trajectories by following $\pi_i^{(t,k)}$ to estimate $\hat{V}_c(\pi^{(t,k)}), \hat{\nabla} V_{\pi_i}^{r_i}(\pi^{(t,k)}), \hat{\nabla} V_{\pi_i}^{c}(\pi^{(t,k)})$

$$\pi_i^{(t,k+1)} = \begin{cases} \mathcal{P}_{\Pi^{i,\xi}} \left[\pi_i^{(t,k)} - \nu_k \hat{\nabla}_{\pi_i} V_{r_i}(\pi^{(t,k)}) - \frac{\nu_k}{\eta} (\pi_i^{(t,k)} - \pi_i^{(t)}) \right] & \text{if } \hat{V}_c(\pi^{(t,k)}) - \alpha \le \delta_k \\ \mathcal{P}_{\Pi^{i,\xi}} \left[\pi_i^{(t,k)} - \nu_k \hat{\nabla}_{\pi_i} V_c(\pi^{(t,k)}) - \frac{\nu_k}{\eta} (\pi_i^{(t,k)} - \pi_i^{(t)}) \right] & \text{otherwise} \end{cases}$$

  $\pi_i^{(t+1)} = \pi_i^{(t,\hat{k})}$ where $\hat{k}$ is sampled from $\{k \in [K] \mid \hat{V}_c(\pi^{(t,k)}) \le \delta_k\}$

output: $\pi_i^{(T)}$ for $i \in \mathcal{N}$

Outline

1. Motivation and Problem Formulation
   - Independent Learning
   - MPGs
   - CMPGs
2. Related Work & Challenges
3. Algorithm
4. Iteration and Sample Complexity Analysis
5. Simulations: Distributed Energy Marketplace

Iteration & Sample Complexity Result

Assumptions

- **Initial feasibility:** $\pi^{(0)}$ *satisfies* $V_c(\pi^{(0)}) < \alpha$

[1] $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic dependencies in $1/\epsilon$, and polynomial dependencies in $m, S, A_{\max}, 1-\gamma, \zeta,\ D$

Iteration & Sample Complexity Result

Assumptions

- **Initial feasibility:** $\pi^{(0)}$ *satisfies* $V_c(\pi^{(0)}) < \alpha$
- **Uniform Slater's condition:**

  $\exists \zeta > 0$ *s.t.* $\forall \pi' \in \Pi$ *with* $V_c(\pi') < \alpha$, $\exists \pi \in \Pi$ *s.t.* $V^c_{\eta,\pi'}(\pi) \leq \alpha - \zeta$

[1] $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic dependencies in $1/\epsilon$, and polynomial dependencies in $m, S, A_{\max}, 1-\gamma, \zeta, D$

Iteration & Sample Complexity Result

Assumptions

- **Initial feasibility:** $\pi^{(0)}$ *satisfies* $V_c(\pi^{(0)}) < \alpha$
- **Uniform Slater's condition:**

$$\exists \zeta > 0 \text{ s.t. } \forall \pi' \in \Pi \text{ with } V_c(\pi') < \alpha,\ \exists \pi \in \Pi \text{ s.t. } V^c_{\eta,\pi'}(\pi) \leq \alpha - \zeta$$

Theorem

For $\epsilon > 0$*, using* ϵ*-greedy exploration, after running iProxCMPG for suitably chosen* η, T, K*, and* $\{(\nu_k, \delta_k)\}_{0 \leq k \leq K}$*,* $\exists t \in [T]$ *s.t. in expectation* $\pi^{(t)}$ *is a constrained* ϵ*-NE.*

[1] $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic dependencies in $1/\epsilon$, and polynomial dependencies in $m, S, A_{\max}, 1-\gamma, \zeta, D$

Iteration & Sample Complexity Result

Assumptions

- **Initial feasibility:** $\pi^{(0)}$ *satisfies* $V_c(\pi^{(0)}) < \alpha$
- **Uniform Slater's condition:**

  $\exists \zeta > 0$ *s.t.* $\forall \pi' \in \Pi$ *with* $V_c(\pi') < \alpha$, $\exists \pi \in \Pi$ *s.t.* $V^c_{\eta,\pi'}(\pi) \le \alpha - \zeta$

Theorem

For $\epsilon > 0$, *using* ϵ*-greedy exploration, after running iProxCMPG for suitably chosen* η, T, K, *and* $\{(\nu_k, \delta_k)\}_{0 \le k \le K}$, $\exists t \in [T]$ *s.t. in expectation* $\pi^{(t)}$ *is a constrained* ϵ*-NE.*

- **Exact gradients:** *total iteration complexity*[1] $\tilde{\mathcal{O}}(\epsilon^{-4})$
- **Finite sample:** *total sample complexity*[1] $\tilde{\mathcal{O}}(\epsilon^{-7})$

---

[1] $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic dependencies in $1/\epsilon$, and polynomial dependencies in $m, S, A_{\max}, 1-\gamma, \zeta, D$

Comparison

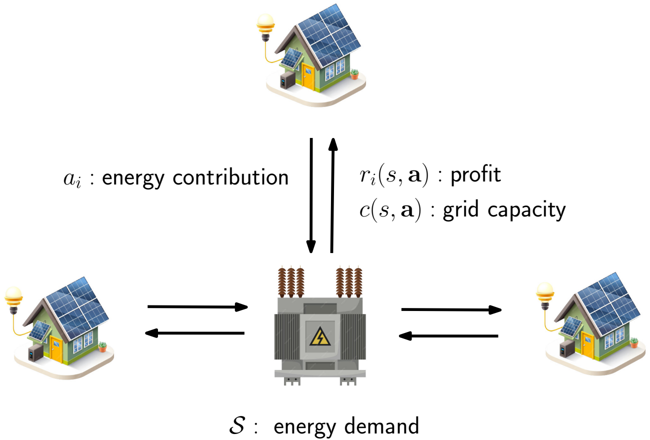
	centralized	**independent**
MPG	Nash-CA [Song et al., 2022] $\mathcal{O}(\epsilon^{-3})$	Independent PGA [Leonardos et al., 2022],[Zhang et al., 2022b] [Ding et al., 2022] $\mathcal{O}(\epsilon^{-5})$
CMPG	CA-CMPG [Alatur et al., 2023] $\tilde{\mathcal{O}}(\epsilon^{-5})$	*Our algorithm* $\tilde{\mathcal{O}}(\epsilon^{-7})$

Outline

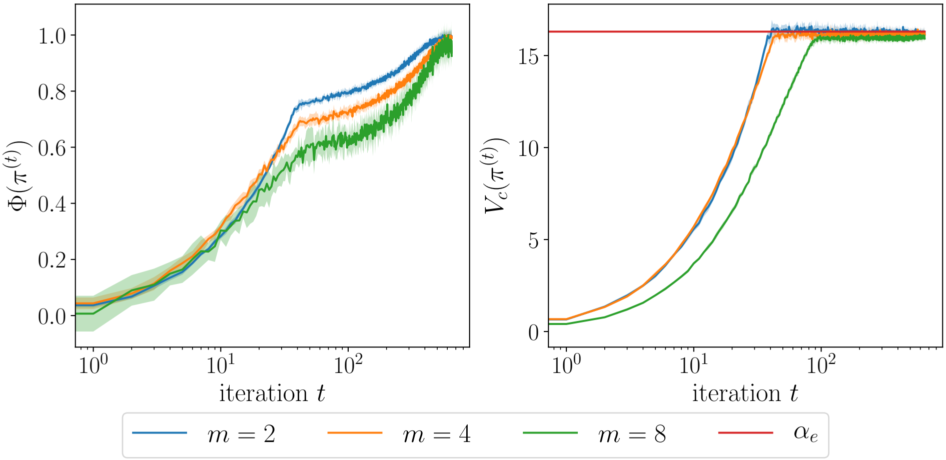
1. Motivation and Problem Formulation
    - Independent Learning
    - MPGs
    - CMPGs
2. Related Work & Challenges
3. Algorithm
4. Iteration and Sample Complexity Analysis
5. Simulations: Distributed Energy Marketplace

Simulations

- **Pollution tax model**
- **Distributed energy marketplace**, inspired by [Narasimha et al., 2022]



Simulations



Future Work

- Sample complexity improvement to match centralized algorithms?
- "Fully" independent learning dynamics (agents with different algorithms)?
- Scaling to large spaces via function approximation
- Beyond CMPGs

References I

Alatur, P., Ramponi, G., He, N., and Krause, A. (2023).
Provably learning nash policies in constrained markov potential games.
In *Sixteenth European Workshop on Reinforcement Learning*.

Altman, E. and Shwartz, A. (2000).
Constrained Markov Games: Nash Equilibria.
In Filar, J. A., Gaitsgory, V., and Mizukami, K., editors, *Advances in Dynamic Games and Applications*, Annals of the International Society of Dynamic Games, pages 213–221, Boston, MA. Birkhäuser.

Boob, D., Deng, Q., and Lan, G. (2023).
Stochastic first-order methods for convex and nonconvex functional constrained optimization.
Mathematical Programming, 197(1):215–279.

References II

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.

Jia, Z. and Grimmer, B. (2023).
First-Order Methods for Nonsmooth Nonconvex Functional Constrained Optimization with or without Slater Points.
arXiv:2212.00927 [math].

References III

Lan, G. and Zhou, Z. (2020).
Algorithms for stochastic optimization with function or expectation constraints.
Computational Optimization and Applications, 76(2):461–498.

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*.

Ma, R., Lin, Q., and Yang, T. (2020).
Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints.
pages 6554–6564. PMLR.

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

References IV

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

Narasimha, D., Lee, K., Kalathil, D., and Shakkottai, S. (2022).
Multi-agent learning via markov potential games in marketplaces for distributed energy resources.
In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6350–6357. IEEE.

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.

References V

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

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.

References VI

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.

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.

Appendix

Simulations

Distributed energy marketplace

- m energy providers, choosing amount of energy to contribute $\mathcal{A}_i = \{0, \ldots, A_i - 1\}$

Simulations

Distributed energy marketplace

- m energy providers, choosing amount of energy to contribute $\mathcal{A}_i = \{0, \ldots, A_i - 1\}$
- $\mathcal{S} = \{0, \ldots, S - 1\}$ energy demand (high to low)

Simulations

Distributed energy marketplace

- m energy providers, choosing amount of energy to contribute $\mathcal{A}_i = \{0, \ldots, A_i - 1\}$
- $\mathcal{S} = \{0, \ldots, S - 1\}$ energy demand (high to low)
- profit $r_i(s, a_i, a_{-i}) = c_0 a_i^2 - c_1 a_i^2 \sum_{i \in \mathcal{N}} a_i - a_i c_2^s$ for some $c_0, c_1, c_2 \in \mathbb{R}$

Simulations

Distributed energy marketplace

- m energy providers, choosing amount of energy to contribute $\mathcal{A}_i = \{0, \ldots, A_i - 1\}$
- $\mathcal{S} = \{0, \ldots, S - 1\}$ energy demand (high to low)
- profit $r_i(s, a_i, a_{-i}) = c_0 a_i^2 - c_1 a_i^2 \sum_{i \in \mathcal{N}} a_i - a_i c_2^s$ for some $c_0, c_1, c_2 \in \mathbb{R}$
- sample $w \sim \mathcal{U}(\{0, 1, \ldots, W\})$ and set

$$s' = \begin{cases} \mathcal{P}_{[0,S-1]} \left(\sum_{i \in \mathcal{N}} a_i - w \right) & \text{w.p. } 0.9 \\ w & \text{w.p. } 0.1 \end{cases}$$

Simulations

Distributed energy marketplace

- m energy providers, choosing amount of energy to contribute $\mathcal{A}_i = \{0, \ldots, A_i - 1\}$
- $\mathcal{S} = \{0, \ldots, S - 1\}$ energy demand (high to low)
- profit $r_i(s, a_i, a_{-i}) = c_0 a_i^2 - c_1 a_i^2 \sum_{i \in \mathcal{N}} a_i - a_i c_2^s$ for some $c_0, c_1, c_2 \in \mathbb{R}$
- sample $w \sim \mathcal{U}(\{0, 1, \ldots, W\})$ and set

$$s' = \begin{cases} \mathcal{P}_{[0,S-1]} \left(\sum_{i \in \mathcal{N}} a_i - w \right) & \text{w.p. } 0.9 \\ w & \text{w.p. } 0.1 \end{cases}$$

- $c(s, \mathbf{a}) = \sum_{i \in \mathcal{N}} a_i$, require $V_c(\pi) \leq \alpha_e$ for some $\alpha_e \in \mathbb{R}$

Simulations

Distributed energy marketplace

- m energy providers, choosing amount of energy to contribute $\mathcal{A}_i = \{0, \dots, A_i - 1\}$
- $\mathcal{S} = \{0, \dots, S - 1\}$ energy demand (high to low)
- profit $r_i(s, a_i, a_{-i}) = c_0 a_i^2 - c_1 a_i^2 \sum_{i \in \mathcal{N}} a_i - a_i c_2^s$ for some $c_0, c_1, c_2 \in \mathbb{R}$
- sample $w \sim \mathcal{U}(\{0, 1, \dots, W\})$ and set

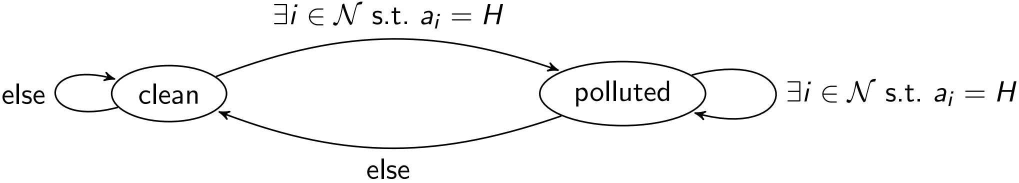
$$s' = \begin{cases} \mathcal{P}_{[0,S-1]}\left(\sum_{i \in \mathcal{N}} a_i - w\right) & \text{w.p. } 0.9 \\ w & \text{w.p. } 0.1 \end{cases}$$

- $c(s, \mathbf{a}) = \sum_{i \in \mathcal{N}} a_i$, require $V_c(\pi) \leq \alpha_e$ for some $\alpha_e \in \mathbb{R}$

$\Longrightarrow$ satisfies CMPG condition, see [Narasimha et al., 2022]

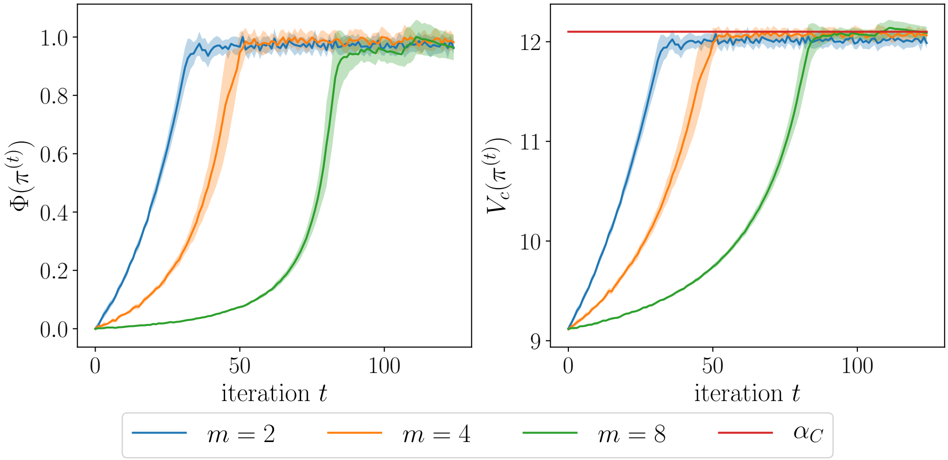
Simulations

Pollution tax model

- m factories that choose production volume $\mathcal{A}_i = \{L, H\}$
- 

- let $r_i(s, \mathbf{a}) = -T_P \mathbb{I}_{\{s=\text{polluted}\}} + \begin{cases} P_L & \text{if } a_i = L \\ P_H & \text{else} \end{cases}$ and $c(s, \mathbf{a}) = |\{i \in \mathcal{N} \mid a_i = H\}|$

Simulations



Iteration & Sample Complexity Result

Assumptions

- **Initial feasibility:** $\pi^{(0)}$ *satisfies* $V_c(\pi^{(0)}) < \alpha$
- **Uniform Slater's condition:**

$$\exists \zeta > 0 \text{ s.t. } \forall \pi' \in \Pi \text{ with } V_c(\pi') < \alpha, \exists \pi \in \Pi \text{ s.t. } V^c_{\eta,\pi'}(\pi) \leq \alpha - \zeta$$

Theorem

For $\epsilon > 0$, *using* ϵ*-greedy exploration, after running iProxCMPG for suitably chosen* η, T, K, *and* $\{(\nu_k, \delta_k)\}_{0 \leq k \leq K}$, *there exists* $t \in [T]$ *s.t. in expectation* $\pi^{(t)}$ *is a constrained* ϵ*-NE.*

- **Exact gradients:** *total iteration complexity*[2] $\tilde{\mathcal{O}}(\epsilon^{-4})$
- **Finite sample:** *total sample complexity*[1] $\tilde{\mathcal{O}}(\epsilon^{-7})$

---

[2] $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic dependencies in $1/\epsilon$, and polynomial dependencies in $m, S, A_{\max}, 1-\gamma, \zeta$, and D.

Proof idea (exact gradients).

1. for $K = \mathcal{O}(\epsilon^{-2})$, inner loop guarantees sufficiently exact proximal update
2. for $T = \mathcal{O}(\epsilon^{-2})$, outer loop guarantees $\exists t \in [T]$ s.t. $\left\|\pi^{(t+1)} - \pi^{(t)}\right\| = \mathcal{O}(\epsilon)$

$\implies$ $\pi^{(t+1)}$ satisfies ϵ-KKT conditions for $\min_{\pi \in \Pi_c} \Phi(\pi)$

$\implies$ $\pi^{(t+1)}$ satisfies ϵ-KKT conditions for playerwise problem with $\pi_{-i}^{(t+1)}$ fixed:

$$\min_{\pi_i \in \Pi_c^i(\pi_{-i}^{(t+1)})} V_{r_i}(\pi_i, \pi_{-i}^{(t+1)}) \tag{2}$$

$\overset{gr.dom.}{\implies}$ for all $i \in \mathcal{N}$, bound duality gap of (2) via gradient dominance

$\implies$ together with $V_c(\pi^{(t+1)}) \leq \alpha$, it follows that $\pi^{(t+1)}$ is constrained $\mathcal{O}(\epsilon)$-NE

$\square$

Proof idea (finite sample).

Algorithm 1 iProxCMPG: **i**ndependent **Prox**imal-policy algorithm for **CMPG**s $\mathcal{O}(\epsilon^{-7})$

1: **initialization:** $\pi^{(0)} \in \Pi^{\xi}$ s.t. $V_c(\pi^{(0)}) < \alpha$ and suitably chosen $\eta, \xi, T, K, \{(\nu_k, \delta_k)\}_{0 \le k \le K}$

2: **for** $t = 0, \dots, T-1$ **do** $\leftarrow$ $\mathcal{O}(\epsilon^{-2})$ times

3: $\quad \pi_i^{(t,0)} = \pi_i^{(t)}$

4: $\quad$ **for** $k = 0, \dots, K-1$ and $i \in \mathcal{N}$ simultaneously **do** $\leftarrow$ $\mathcal{O}(\epsilon^{-3})$ times

5: $\quad\quad$ sample B trajectories by following $\pi_i^{(t,k)}$ to estimate $\hat{V}_{r_i}(\pi^{(t,k)})$, $\underline{\hat{\nabla} V^{r_i}_{\pi_i}(\pi^{(t,k)}), \hat{\nabla} V^{c}_{\pi_i}(\pi^{(t,k)})}$

$B = \mathcal{O}(\epsilon^{-2})$

6: $$\pi_i^{(t,k+1)} = \begin{cases} \mathcal{P}_{\Pi^{i,\xi}}\left[\pi_i^{(t,k)} - \nu_k \hat{\nabla}_{\pi_i} V_{r_i}(\pi^{(t,k)}) - \frac{\nu_k}{\eta}(\pi_i^{(t,k)} - \pi_i^{(t)})\right] & \text{if } \hat{V}_c(\pi^{(t,k)}) - \alpha \le \delta_k \\ \mathcal{P}_{\Pi^{i,\xi}}\left[\pi_i^{(t,k)} - \nu_k \hat{\nabla}_{\pi_i} V_c(\pi^{(t,k)}) - \frac{\nu_k}{\eta}(\pi_i^{(t,k)} - \pi_i^{(t)})\right] & \text{otherwise} \end{cases}$$

7: $\quad \pi_i^{(t+1)} = \pi_i^{(t,\hat{k})}$ where $\hat{k}$ is sampled from $\{k \in [K] \mid \hat{V}_c(\pi^{(t,k)}) \le \delta_k\}$

variance $\mathcal{O}(\epsilon^{-1})$

8: **output:** $\pi_i^{(T)}$ for $i \in \mathcal{N}$

$\square$

Approach 1: Primal-dual method

$$\mathcal{L}(\pi, \lambda) = \Phi(\pi) + \lambda(V_c(\pi) - \alpha)$$

- if strong duality holds, then

$$\inf_{\pi \in \Pi} \sup_{\lambda \geq 0} \mathcal{L}(\pi, \lambda) = \sup_{\lambda \geq 0} \inf_{\pi \in \Pi} \mathcal{L}(\pi, \lambda)$$

Approach 1: Primal-dual method

$$\mathcal{L}(\pi, \lambda) = \Phi(\pi) + \lambda(V_c(\pi) - \alpha)$$

- if ~~strong duality~~ [Alatur et al., 2023] holds, then

$$\inf_{\pi \in \Pi} \sup_{\lambda \geq 0} \mathcal{L}(\pi, \lambda) = \sup_{\lambda \geq 0} \inf_{\pi \in \Pi} \mathcal{L}(\pi, \lambda)$$