A Gentle Introduction to Multi-Agent Reinforcement Learning

Kaiqing Zhang

EPFL-ETH Summer School

July 30, 2024

Disclaimers...

- No single-agent reinforcement learning (RL) nor Game Theory basics (see wonderful tutorials yesterday!)

Disclaimers...

- No single-agent reinforcement learning (RL) nor Game Theory basics (see wonderful tutorials yesterday!)
- No fancy videos/demos, nor “deep” neural nets/“Transformers”

Disclaimers...

- No single-agent reinforcement learning (RL) nor Game Theory basics (see wonderful tutorials yesterday!)
- No fancy videos/demos, nor “deep” neural nets/“Transformers”
- No (GitHub) codebases or Jupyter notebook, PyTorch, PettingZoo (Terry et al., 2021), OpenSpiel (Lanctot et al., 2019), StarCraft Multi-Agent Challenge (SMAC) (Samvelyan et al., 2019), or Google Football (Kurach et al., 2020)...

Disclaimers...

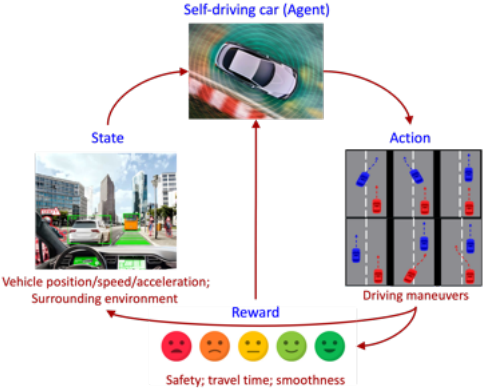
- No single-agent reinforcement learning (RL) nor Game Theory basics (see wonderful tutorials yesterday!)
- No fancy videos/demos, nor "deep" neural nets/"Transformers"
- No (GitHub) codebases or Jupyter notebook, PyTorch, PettingZoo (Terry et al., 2021), OpenSpiel (Lanctot et al., 2019), StarCraft Multi-Agent Challenge (SMAC) (Samvelyan et al., 2019), or Google Football (Kurach et al., 2020)...
- May not be most comprehensive and up-to-date (but will try :))

Reinforcement Learning

- Reinforcement learning (RL) has attracted increasing attention lately

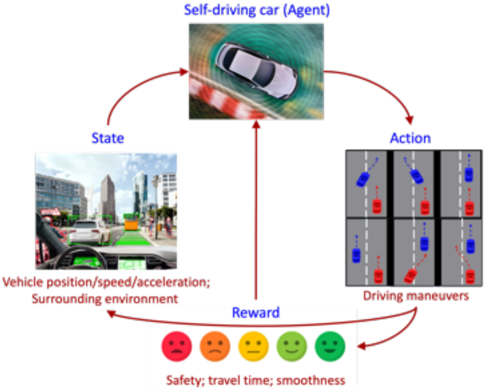
Reinforcement Learning

- Reinforcement learning (RL) has attracted increasing attention lately
- Goal: Autonomous agents make sequential decisions in unknown dynamic environments



Reinforcement Learning

- Reinforcement learning (RL) has attracted increasing attention lately
- Goal: Autonomous agents make sequential decisions in unknown dynamic environments



Multi-agent Interactions are Prevalent in AI Systems

- In fact, many success stories of AI systems naturally involve multi-agent interactions in a dynamic environment:

Multi-agent Interactions are Prevalent in AI Systems

- In fact, many success stories of AI systems naturally involve multi-agent interactions in a dynamic environment:





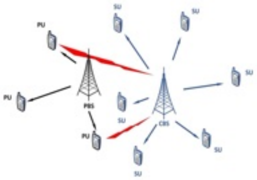

- Examples (left-to-right): Self-driving fleets (Shalev-Shwartz et al., 2016), networked robotic arms (Levine et al., 2018), Amazon warehouse robots (Amazon, 2023), DeepBlue (IBM, 1997), AlphaGo (Silver et al., 2016), poker bots (Heinrich and Silver, 2016)

Multi-agent Interactions are Prevalent in AI Systems

- And many more (you must have heard of)... Dota 5v5 (Berner et al., 2019), Diplomacy games (FAIR, 2022), StarCraft games (Vinyals et al., 2019), particle environments (Lowe et al., 2017)

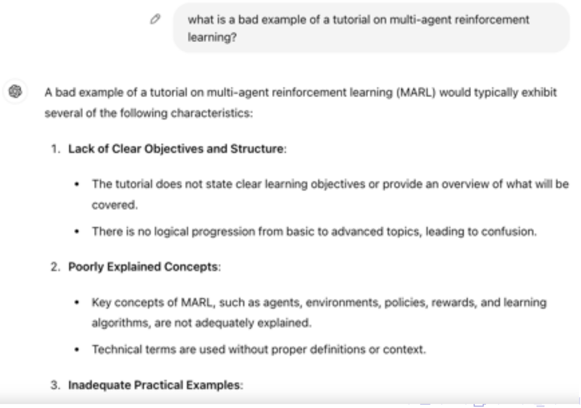
Multi-agent Interactions are Prevalent in AI Systems

- And many more (you might have heard of)... telecommunications (Liu and Zhao, 2010; Bubeck et al., 2020), finance/trading (Liu et al., 2022c), power systems (Chen et al., 2022a)



Multi-agent Interactions are Prevalent in AI Systems

- And many more (you/I may not have thought of)... "field deployment" of variable speed limit control at Nashville, Tennessee (Zhang et al., 2024b); ChatBot and conversational agent (OpenAI, 2022)



Multi-agent Reinforcement Learning

Multi-agent Reinforcement Learning

- An old topic that got revitalized recently, probably all started from

Multi-agent Reinforcement Learning

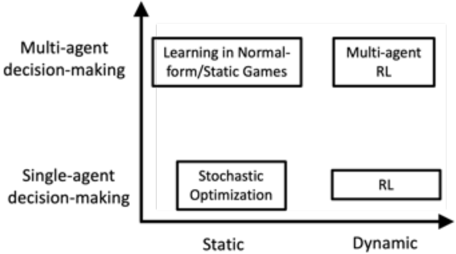
- An old topic that got revitalized recently, probably all started from

- Received broad research interest from ML, Econ, Control, and Alg. Game Theory (with an increasing number of workshops/programs at Simons Institute, NeurIPS, ICML, ICLR, CDC ... over the years)

Multi-agent Reinforcement Learning

- An old topic that got revitalized recently, probably all started from

- Received broad research interest from ML, Econ, Control, and Alg. Game Theory (with an increasing number of workshops/programs at Simons Institute, NeurIPS, ICML, ICLR, CDC ... over the years)
  - What is really multi-agent RL (MARL)? In one figure:



A Gentle Introduction to MARL: Outline

- Part I: Basics and Classical Results
- Part II: Modern Results
- Part III: Why Multi-agent RL?
- Concluding Remarks

Part I.A: Basics

A Basic Model: Stochastic/Markov Games (SGs/MGs)

A Basic Model: Stochastic/Markov Games (SGs/MGs)

- (Infinite-horizon) stochastic games (Shapley, 1953; Fink et al., 1964):
  $$\langle S, \{A^i\}_{i\in[n]}, \{r_s^i\}_{s\in S, i\in[n]}, p, \gamma, \rho\rangle$$
- n agents (called interchangeably as players)
- S is the set of states
- A^i is the set of actions that player i can take
- $r_s^i(a^1, \cdots, a^n)$ is reward of player i given joint action $(a^1, \cdots, a^n)$ at s;
  - If $n = 2$ and $r_s^1(a^1, a^2) + r_s^2(a^1, a^2) = 0$, it is two-player zero-sum; competitive nature
  - If $r^1 = r^2 = \cdots = r^n$, it is identical-interest or common-payoff or a team problem; cooperative nature
- Player i takes actions $a^i \in A^i$ at state $s \in S$, and the state transitions to s' according to $s' \sim p(\cdot|s, a^1, \cdots, a^n) \in \Delta(S)$
- $\gamma \in [0, 1)$ is the discount factor; $\rho \in \Delta(\mathcal{S})$ is the initial state distribution

A Basic Model: Stochastic/Markov Games (SGs/MGs)

- (Infinite-horizon) stochastic games (Shapley, 1953; Fink et al., 1964):
$$\langle S, \{A^i\}_{i\in[n]}, \{r_s^i\}_{s\in S, i\in[n]}, p, \gamma, \rho\rangle$$
- n agents (called interchangeably as players)
- S is the set of states
- A^i is the set of actions that player i can take
- $r_s^i(a^1, \cdots, a^n)$ is reward of player i given joint action $(a^1, \cdots, a^n)$ at s;
  - If $n = 2$ and $r_s^1(a^1, a^2) + r_s^2(a^1, a^2) = 0$, it is two-player zero-sum; competitive nature
  - If $r^1 = r^2 = \cdots = r^n$, it is identical-interest or common-payoff or a team problem; cooperative nature
- Player i takes actions $a^i \in A^i$ at state $s \in S$, and the state transitions to s' according to $s' \sim p(\cdot|s, a^1, \cdots, a^n) \in \Delta(S)$
- $\gamma \in [0, 1)$ is the discount factor; $\rho \in \Delta(\mathcal{S})$ is the initial state distribution
- As a fundamental framework for MARL ever since (Littman, 1994)

A Basic Model: Stochastic/Markov Games

- Finite-horizon/Episodic variant (common in recent MARL theory):
$$\langle S, \{A^i\}_{i\in[n]}, \{r_s^{i,h}\}_{s\in S, i\in[n], h\in[H]}, \{p^h\}_{h\in[H]}, H\rangle$$
- S is the set of states
- A^i is the set of actions that player i can take
- $r_s^{i,h}(a^1, \cdots, a^n)$ denotes the reward function of player i given action profile $(a^1, \cdots, a^n)$ at state s and step h;
- Player i takes actions $a_h^i \in A^i$ at state $s_h \in S$ and step h, and the state transitions to s_{h+1} at $h+1$ by $s_{h+1} \sim p^h(\cdot|s_h, a_h^1, \cdots, a_h^n) \in \Delta(S)$
- H is the episode length

Infinite-horizon SGs: Policies

- Mostly consider stationary Markov policies (as usual in single-agent RL)
- Let $\pi^i := \{\pi^i(s)\}_{s \in S}$ with $\pi^i(s)$ (or π^i_s for short) in $\Delta(\mathcal{A}^i)$ denoting the (mixed) strategy of player i at state s and $\pi = (\pi^1, \cdots, \pi^n)$ denoting a joint policy

Infinite-horizon SGs: Policies

- Mostly consider stationary Markov policies (as usual in single-agent RL)
- Let $\pi^i := \{\pi^i(s)\}_{s \in S}$ with $\pi^i(s)$ (or π^i_s for short) in $\Delta(\mathcal{A}^i)$ denoting the (mixed) strategy of player i at state s and $\pi = (\pi^1, \cdots, \pi^n)$ denoting a joint policy
- One can also define non-stationary Markov policies: $\pi^i = (\pi^{i,1}, \pi^{i,2}, \cdots)$ with $\pi^{i,h}(s)$ (or $\pi^{i,h}_s$) in $\Delta(\mathcal{A}^i)$ at time step h

Infinite-horizon SGs: Policies

- Mostly consider stationary Markov policies (as usual in single-agent RL)
- Let $\pi^i := \{\pi^i(s)\}_{s \in S}$ with $\pi^i(s)$ (or π^i_s for short) in $\Delta(\mathcal{A}^i)$ denoting the (mixed) strategy of player i at state s and $\pi = (\pi^1, \cdots, \pi^n)$ denoting a joint policy
- One can also define non-stationary Markov policies: $\pi^i = (\pi^{i,1}, \pi^{i,2}, \cdots)$ with $\pi^{i,h}(s)$ (or $\pi^{i,h}_s$) in $\Delta(\mathcal{A}^i)$ at time step h
- Joint Markov policies:
  - Stationary: $\pi : \mathcal{S} \to \Delta(\prod_{i=1}^n \mathcal{A}^i)$;
  - Non-stationary: $\pi = (\pi^1, \pi^2, \cdots)$ with $\pi^h : \mathcal{S} \to \Delta(\prod_{i=1}^n \mathcal{A}^i)$ at time step h
- Product policies: $\pi_s = \pi^1_s \times \cdots \times \pi^n_s$, i.e., no correlation in action choice among agents at each state s; otherwise they are correlated in general

Infinite-horizon SGs: Policies

- Mostly consider stationary Markov policies (as usual in single-agent RL)
- Let $\pi^i := \{\pi^i(s)\}_{s\in S}$ with $\pi^i(s)$ (or π^i_s for short) in $\Delta(\mathcal{A}^i)$ denoting the (mixed) strategy of player i at state s and $\pi = (\pi^1, \cdots, \pi^n)$ denoting a joint policy
- One can also define non-stationary Markov policies: $\pi^i = (\pi^{i,1}, \pi^{i,2}, \cdots)$ with $\pi^{i,h}(s)$ (or $\pi^{i,h}_s$) in $\Delta(\mathcal{A}^i)$ at time step h
- Joint Markov policies:
  - Stationary: $\pi : \mathcal{S} \to \Delta(\prod_{i=1}^n \mathcal{A}^i)$;
  - Non-stationary: $\pi = (\pi^1, \pi^2, \cdots)$ with $\pi^h : \mathcal{S} \to \Delta(\prod_{i=1}^n \mathcal{A}^i)$ at time step h
- Product policies: $\pi_s = \pi^1_s \times \cdots \times \pi^n_s$, i.e., no correlation in action choice among agents at each state s; otherwise they are correlated in general
- Marginalized policies of other agents $-i$: given π and agent i, $\pi^{-i} : \mathcal{S} \to \Delta(\mathcal{A}^{-i})$ outputs its marginal distribution at each state s

Infinite-horizon SGs: Policies

- Mostly consider stationary Markov policies (as usual in single-agent RL)
- Let $\pi^i := \{\pi^i(s)\}_{s\in S}$ with $\pi^i(s)$ (or π^i_s for short) in $\Delta(\mathcal{A}^i)$ denoting the (mixed) strategy of player i at state s and $\pi = (\pi^1, \cdots, \pi^n)$ denoting a joint policy
- One can also define non-stationary Markov policies: $\pi^i = (\pi^{i,1}, \pi^{i,2}, \cdots)$ with $\pi^{i,h}(s)$ (or $\pi^{i,h}_s$) in $\Delta(\mathcal{A}^i)$ at time step h
- Joint Markov policies:
    - Stationary: $\pi : \mathcal{S} \to \Delta(\prod_{i=1}^n \mathcal{A}^i)$;
    - Non-stationary: $\pi = (\pi^1, \pi^2, \cdots)$ with $\pi^h : \mathcal{S} \to \Delta(\prod_{i=1}^n \mathcal{A}^i)$ at time step h
- Product policies: $\pi_s = \pi^1_s \times \cdots \times \pi^n_s$, i.e., no correlation in action choice among agents at each state s; otherwise they are correlated in general
- Marginalized policies of other agents $-i$: given π and agent i, $\pi^{-i} : \mathcal{S} \to \Delta(\mathcal{A}^{-i})$ outputs its marginal distribution at each state s
- Will focus on Markov policies throughout unless otherwise noted

Infinite-horizon SGs: Value Functions and Best-responses

- Define the state-value function of player i as

$$V_\pi^i(s) := \mathbb{E}_{a_k \sim \pi_{s_k}} \left\{ \sum_{k=0}^{\infty} \gamma^k r_{s_k}^i(a_k) \middle| s_0 = s \right\}, \forall s$$

  where $\{s_k\}_{k \geq 0}$ is a state process under joint policy π

- Other (state-action-)value functions may be useful:

$$Q_\pi^i(s,a) := \mathbb{E}_{a_k \sim \pi_{s_k}} \left\{ \sum_{k=0}^{\infty} \gamma^k r_{s_k}^i(a_k) \middle| s_0 = s, a_0 = a \right\}, \forall s, a$$

$$q_\pi^i(s,a^i) := \mathbb{E}_{a_k^{-i} \sim \pi_{s_k}^{-i}}[Q_\pi^i(s,a^i,a^{-i})]$$

Infinite-horizon SGs: Value Functions and Best-responses

- Define the state-value function of player i as

$$V^i_\pi(s) := \mathbb{E}_{a_k\sim\pi_{s_k}}\left\{\sum_{k=0}^{\infty}\gamma^k r^i_{s_k}(a_k)\middle| s_0 = s\right\}, \forall s$$

where $\{s_k\}_{k\geq 0}$ is a state process under joint policy π

- Other (state-action-)value functions may be useful:

$$Q^i_\pi(s,a) := \mathbb{E}_{a_k\sim\pi_{s_k}}\left\{\sum_{k=0}^{\infty}\gamma^k r^i_{s_k}(a_k)\middle| s_0 = s, a_0 = a\right\}, \forall s, a$$

$$q^i_\pi(s,a^i) := \mathbb{E}_{a_k^{-i}\sim\pi_{s_k}^{-i}}[Q^i_\pi(s,a^i,a^{-i})]$$

- Best-response policies: for a stationary policy $\pi^{-i}:\mathcal{S}\to\Delta(\mathcal{A}^{-i})$, the best-response policy of agent i is $\pi^i_\dagger(\pi^{-i})$ such that

$$V^i_{\dagger,\pi^{-i}}(s) := V^i_{\pi^i_\dagger(\pi^{-i})\times\pi^{-i}}(s) = \max_{\widetilde{\pi}^i:\mathcal{S}\to\Delta(\mathcal{A}^i)} V^i_{\widetilde{\pi}^i\times\pi^{-i}}(s)$$

Infinite-horizon SGs: Value Functions and Best-responses

- Define the state-value function of player i as

$$V^i_\pi(s) := \mathbb{E}_{a_k \sim \pi_{s_k}} \left\{ \sum_{k=0}^{\infty} \gamma^k r^i_{s_k}(a_k) \middle| s_0 = s \right\}, \forall s$$

  where $\{s_k\}_{k\geq 0}$ is a state process under joint policy π

- Other (state-action-)value functions may be useful:

$$Q^i_\pi(s,a) := \mathbb{E}_{a_k \sim \pi_{s_k}} \left\{ \sum_{k=0}^{\infty} \gamma^k r^i_{s_k}(a_k) \middle| s_0 = s, a_0 = a \right\}, \forall s, a$$

$$q^i_\pi(s,a^i) := \mathbb{E}_{a_k^{-i} \sim \pi_{s_k}^{-i}}[Q^i_\pi(s,a^i,a^{-i})]$$

- Best-response policies: for a stationary policy $\pi^{-i} : \mathcal{S} \to \Delta(\mathcal{A}^{-i})$, the best-response policy of agent i is $\pi^i_\dagger(\pi^{-i})$ such that

$$V^i_{\dagger,\pi^{-i}}(s) := V^i_{\pi^i_\dagger(\pi^{-i})\times\pi^{-i}}(s) = \max_{\widetilde{\pi}^i : \mathcal{S}\to\Delta(\mathcal{A}^i)} V^i_{\widetilde{\pi}^i \times \pi^{-i}}(s)$$

- Since π^{-i} is Markov, there exists a $\pi^i_\dagger(\pi^{-i})$ that best-responding at all s (essentially an MDP from agent i's perspective)

Infinite-horizon SGs: Solution Concepts

- Strategy modification: $\phi^i : \mathcal{S} \times \mathcal{A}^i \to \mathcal{A}^i$ can modify the action of agent i, after seeing the action recommended by π; denote the modified joint policy as $(\phi^i \diamond \pi^i) \odot \pi^{-i}$
  - Different strategy modification classes exist, e.g., history-dependent

Infinite-horizon SGs: Solution Concepts

- Strategy modification: $\phi^i : \mathcal{S} \times \mathcal{A}^i \to \mathcal{A}^i$ can modify the action of agent i, after seeing the action recommended by π; denote the modified joint policy as $(\phi^i \diamond \pi^i) \odot \pi^{-i}$
  - Different strategy modification classes exist, e.g., history-dependent
- Common solution concepts:

Definition ((Markov Perfect Stationary) Nash Equilibrium)

A joint product Markov stationary policy $\pi_* = (\pi_*^1, \cdots, \pi_*^n)$ is an ϵ-(Markov perfect stationary) Nash-equilibrium (NE) provided that

$$\texttt{NE-Gap}(\pi_*) := \max_{i \in [n], s \in \mathcal{S}} \left\{ \max_{\widetilde{\pi}^i : \mathcal{S} \to \Delta(\mathcal{A}^i)} V^i_{\widetilde{\pi}^i \times \pi_*^{-i}}(s) - V^i_{\pi_*}(s) \right\} \leq \epsilon,$$

with $\epsilon = 0$ corresponding to the (Markov perfect stationary) NE.

Infinite-horizon SGs: Solution Concepts

- Strategy modification: $\phi^i : \mathcal{S} \times \mathcal{A}^i \to \mathcal{A}^i$ can modify the action of agent i, after seeing the action recommended by π; denote the modified joint policy as $(\phi^i \diamond \pi^i) \odot \pi^{-i}$
  - Different strategy modification classes exist, e.g., history-dependent
- Common solution concepts:

Definition ((Markov Perfect Stationary) Nash Equilibrium)

A joint product Markov stationary policy $\pi_* = (\pi_*^1, \cdots, \pi_*^n)$ is an ϵ-(Markov perfect stationary) Nash-equilibrium (NE) provided that

$$\texttt{NE-Gap}(\pi_*) := \max_{i \in [n], s \in \mathcal{S}} \left\{ \max_{\widetilde{\pi}^i : \mathcal{S} \to \Delta(\mathcal{A}^i)} V^i_{\widetilde{\pi}^i \times \pi_*^{-i}}(s) - V^i_{\pi_*}(s) \right\} \leq \epsilon,$$

with $\epsilon = 0$ corresponding to the (Markov perfect stationary) NE.

- Always exists for finite-space SGs (Shapley, 1953; Fink et al., 1964)

Infinite-horizon SGs: Solution Concepts

Definition ((Markov Perfect Stationary) Coarse Correlated Equilibrium)

A joint Markov stationary policy $\pi_* = (\pi_*^1, \cdots, \pi_*^n)$ is an ϵ-(Markov perfect stationary) coarse correlated equilibrium (CCE) provided that

$$\texttt{CCE-Gap}(\pi_*) := \max_{i\in[n], s\in\mathcal{S}} \left\{ \max_{\widetilde{\pi}^i:\mathcal{S}\to\Delta(\mathcal{A}^i)} V^i_{\widetilde{\pi}^i\times\pi_*^{-i}}(s) - V^i_{\pi_*}(s) \right\} \le \epsilon,$$

with $\epsilon = 0$ corresponding to the (Markov perfect stationary) CCE.

Infinite-horizon SGs: Solution Concepts

Definition ((Markov Perfect Stationary) Coarse Correlated Equilibrium)

A joint Markov stationary policy $\pi_* = (\pi_*^1, \cdots, \pi_*^n)$ is an ϵ-(Markov perfect stationary) coarse correlated equilibrium (CCE) provided that

$$\texttt{CCE-Gap}(\pi_*) := \max_{i\in[n], s\in\mathcal{S}} \left\{ \max_{\widetilde{\pi}^i:\mathcal{S}\to\Delta(\mathcal{A}^i)} V^i_{\widetilde{\pi}^i\times\pi_*^{-i}}(s) - V^i_{\pi_*}(s) \right\} \leq \epsilon,$$

with $\epsilon = 0$ corresponding to the (Markov perfect stationary) CCE.

Definition ((Markov Perfect Stationary) Correlated Equilibrium)

A joint Markov stationary policy $\pi_* = (\pi_*^1, \cdots, \pi_*^n)$ is an ϵ-(Markov perfect stationary) correlated equilibrium (CE) provided that

$$\texttt{CE-Gap}(\pi_*) := \max_{i\in[n], s\in\mathcal{S}} \left\{ \max_{\phi^i} V^i_{(\phi^i\diamond\pi_*^i)\odot\pi_*^{-i}}(s) - V^i_{\pi_*}(s) \right\} \leq \epsilon,$$

with $\epsilon = 0$ corresponding to the (Markov perfect stationary) CE.

- Also exist due to NE $\subseteq$ CE $\subseteq$ CCE
- Can define non-stationary versions of the equilibria correspondingly

Finite-horizon SGs: Policies, Values, Solution Concepts

- Should consider non-stationary policies: for each agent i, $\pi^i = (\pi^{i,1}, \cdots, \pi^{i,H})$ with $\pi_s^{i,h} \in \Delta(\mathcal{A}^i)$ at step h
- State-value function (for step $h \in [H]$):

$$V_\pi^{i,h}(s_h) := \mathbb{E}_{a_{h'} \sim \pi_{s_{h'}}} \left\{ \sum_{h'=h}^{H} r_{s_{h'}}^{i,h'}(a_{h'}) \Big| s_h \right\},$$

Finite-horizon SGs: Policies, Values, Solution Concepts

- Should consider non-stationary policies: for each agent i, $\pi^i = (\pi^{i,1}, \cdots, \pi^{i,H})$ with $\pi_s^{i,h} \in \Delta(\mathcal{A}^i)$ at step h
- State-value function (for step $h \in [H]$):

$$V_\pi^{i,h}(s_h) := \mathbb{E}_{a_{h'} \sim \pi_{s_{h'}}} \left\{ \sum_{h'=h}^{H} r_{s_{h'}}^{i,h'}(a_{h'}) \Big| s_h \right\},$$

- Best-responses, strategy modifications, and NE, CE, CCE are oftentimes defined with respect to $V_\pi^{i,1}(s_1)$ at time step 1, e.g., for ϵ-NE

$$\texttt{NE-Gap}(\pi_*) := \max_{i \in [n]} \left\{ V_{\dagger, \pi_*^{-i}}^{i,1}(s_1) - V_{\pi_*}^{i,1}(s_1) \right\} \leq \epsilon$$

Finite-horizon SGs: Policies, Values, Solution Concepts

- Should consider non-stationary policies: for each agent i, $\pi^i = (\pi^{i,1}, \cdots, \pi^{i,H})$ with $\pi_s^{i,h} \in \Delta(\mathcal{A}^i)$ at step h
- State-value function (for step $h \in [H]$):

$$V_\pi^{i,h}(s_h) := \mathbb{E}_{a_{h'} \sim \pi_{s_{h'}}} \left\{ \sum_{h'=h}^{H} r_{s_{h'}}^{i,h'}(a_{h'}) \Big| s_h \right\},$$

- Best-responses, strategy modifications, and NE, CE, CCE are oftentimes defined with respect to $V_\pi^{i,1}(s_1)$ at time step 1, e.g., for ϵ-NE

$$\texttt{NE-Gap}(\pi_*) := \max_{i \in [n]} \left\{ V_{\dagger,\pi_*^{-i}}^{i,1}(s_1) - V_{\pi_*}^{i,1}(s_1) \right\} \leq \epsilon$$

- With $H = \mathcal{O}\left(\frac{\log(1/\epsilon)}{1-\gamma}\right)$, the non-stationary solution concepts in both cases become $\mathcal{O}(\epsilon)$-close
  - Can use finite-horizon algorithms to find approximate non-stationary solution for infinite-horizon settings

Finite-horizon SGs: Policies, Values, Solution Concepts

- Should consider non-stationary policies: for each agent i, $\pi^i = (\pi^{i,1}, \cdots, \pi^{i,H})$ with $\pi_s^{i,h} \in \Delta(\mathcal{A}^i)$ at step h
- State-value function (for step $h \in [H]$):

$$V_\pi^{i,h}(s_h) := \mathbb{E}_{a_{h'} \sim \pi_{s_{h'}}} \left\{ \sum_{h'=h}^{H} r_{s_{h'}}^{i,h'}(a_{h'}) \Big| s_h \right\},$$

- Best-responses, strategy modifications, and NE, CE, CCE are oftentimes defined with respect to $V_\pi^{i,1}(s_1)$ at time step 1, e.g., for ϵ-NE

$$\texttt{NE-Gap}(\pi_*) := \max_{i \in [n]} \left\{ V_{\dagger, \pi_*^{-i}}^{i,1}(s_1) - V_{\pi_*}^{i,1}(s_1) \right\} \leq \epsilon$$

- With $H = \mathcal{O}\left(\frac{\log(1/\epsilon)}{1-\gamma}\right)$, the non-stationary solution concepts in both cases become $\mathcal{O}(\epsilon)$-close
  - Can use finite-horizon algorithms to find approximate non-stationary solution for infinite-horizon settings
  - Also works for approximating stationary solution in certain games (come back later)

Planning: Solution Computation with Model knowledge

Planning: Solution Computation with Model knowledge

- Recall the three approaches from single-agent MDPs/RL: value iteration (VI), policy iteration (PI), and linear programming (LP)

Planning: Solution Computation with Model knowledge

- Recall the three approaches from single-agent MDPs/RL: value iteration (VI), policy iteration (PI), and linear programming (LP)
- Value iteration: let $\boldsymbol{V}_*^h := (V_*^{1,h}, \cdots, V_*^{n,h})$ and $\boldsymbol{r}^h := (\boldsymbol{r}^{1,h}, \cdots, \boldsymbol{r}^{n,h})$

$$\boldsymbol{V}_*^h = \mathcal{B}^h(\boldsymbol{V}_*^{h+1}) := \texttt{Equilibrium}\left[\boldsymbol{r}^h + \gamma \cdot p^h\left[\boldsymbol{V}_*^{h+1}\right]\right], \quad \text{or}$$

$$V_*^{i,h}(s_h) = r^{i,h}(s_h, \pi_*^h) + \gamma \cdot \sum_{s_{h+1}} p^h(s_{h+1} \mid s_h, \pi_*^h) V_*^{i,h+1}(s_{h+1})$$

where π_*^h is the output from some matrix-game equilibrium computation oracle `Equilibrium`, and $\mathcal{B}^h$ is the Bellman operator for SGs

  - Finite-horizon: $\gamma = 1$, $V_*^{i,H+1}(s) = 0$ for all i, s; stops in H-steps
  - Infinite-horizon: $\gamma < 1$, $\boldsymbol{r}^h = \boldsymbol{r}$, $p^h = p$, and thus $\mathcal{B}^h = \mathcal{B}$ for all h

Planning: Value Iteration

- Example: Two-player zero-sum SGs (Shapley, 1953)
  - Minimax Theorem holds:
  $$\max_{\pi^1}\min_{\pi^2} V^h_{\pi^1\times\pi^2} = \min_{\pi^2}\max_{\pi^1} V^h_{\pi^1\times\pi^2},$$
  thus a unique NE value V^h_* for each h
  - CCE collapses to NE

Planning: Value Iteration

- Example: Two-player zero-sum SGs (Shapley, 1953)
  - Minimax Theorem holds:

$$\max_{\pi^1} \min_{\pi^2} V^h_{\pi^1 \times \pi^2} = \min_{\pi^2} \max_{\pi^1} V^h_{\pi^1 \times \pi^2},$$

    thus a unique NE value V^h_* for each h
  - CCE collapses to NE
  - In this case, (minimax) VI proceeds as follows:

$$V^h_*(s) \leftarrow \underbrace{\max_{\mu \in \Delta(\mathcal{A}^1)} \min_{\nu \in \Delta(\mathcal{A}^2)} \mathbb{E}_{a^1 \sim \mu, a^2 \sim \nu}}_{\texttt{Equilibrium oracle}} \left[Q^h_*(s, a^1, a^2) \right]$$

    where matrix $Q^h_*(s, \cdot, \cdot)$

$$Q^h_*(s, a^1, a^2) := r^h(s, a^1, a^2) + \gamma \cdot \sum_{s'} p^h(s' \mid s, a^1, a^2) V^{h+1}_*(s')$$

Planning: Value Iteration

- Example: Two-player zero-sum SGs (Shapley, 1953)
  - Minimax Theorem holds:
$$\max_{\pi^1}\min_{\pi^2} V^h_{\pi^1\times\pi^2} = \min_{\pi^2}\max_{\pi^1} V^h_{\pi^1\times\pi^2},$$
    thus a unique NE value V^h_* for each h
  - CCE collapses to NE
  - In this case, (minimax) VI proceeds as follows:
$$V^h_*(s) \leftarrow \underbrace{\max_{\mu\in\Delta(\mathcal{A}^1)}\min_{\nu\in\Delta(\mathcal{A}^2)} \mathbb{E}_{a^1\sim\mu, a^2\sim\nu}}_{\texttt{Equilibrium oracle}} \left[Q^h_*(s, a^1, a^2)\right]$$
    where matrix $Q^h_*(s,\cdot,\cdot)$
$$Q^h_*(s, a^1, a^2) := r^h(s, a^1, a^2) + \gamma \cdot \sum_{s'} p^h(s' \mid s, a^1, a^2) V^{h+1}_*(s')$$
  - Can also define VI for Q-function (will be used later)
$$Q^h_*(s, a^1, a^2) \leftarrow r^h(s, a^1, a^2) + \gamma \cdot \sum_{s'} p^h(s' \mid s, a^1, a^2)$$
$$\cdot \max_{\mu\in\Delta(\mathcal{A}^1)}\min_{\nu\in\Delta(\mathcal{A}^2)} \mathbb{E}_{\widetilde{a}^1\sim\mu, \widetilde{a}^2\sim\nu}\left[Q^{h+1}_*(s', \widetilde{a}^1, \widetilde{a}^2)\right]$$

Planning: Value Iteration

- Example: Two-player zero-sum SGs (Shapley, 1953)
  - For infinite-horizon case, $\mathcal{B}$ is γ-contracting:

$$\|\mathcal{B}(V) - \mathcal{B}(\widetilde{V})\|_\infty \leq \gamma \cdot \|V - \widetilde{V}\|_\infty$$

Planning: Value Iteration

- Example: Two-player zero-sum SGs (Shapley, 1953)
  - For infinite-horizon case, $\mathcal{B}$ is γ-contracting:
$$\|\mathcal{B}(V) - \mathcal{B}(\widetilde{V})\|_\infty \leq \gamma \cdot \|V - \widetilde{V}\|_\infty$$
  - Key: non-expansiveness of max min operator. For all $s \in \mathcal{S}$

$$V(s) - \widetilde{V}(s)$$
$$= \left| \max_{\mu \in \Delta(\mathcal{A}^1)} \min_{\nu \in \Delta(\mathcal{A}^2)} \mathbb{E}_{a^1 \sim \mu, a^2 \sim \nu} \left[Q(s, a^1, a^2) \right] - \max_{\mu \in \Delta(\mathcal{A}^1)} \min_{\nu \in \Delta(\mathcal{A}^2)} \mathbb{E}_{a^1 \sim \mu, a^2 \sim \nu} \left[\widetilde{Q}(s, a^1, a^2) \right] \right|$$
$$\leq \|Q(s, \cdot, \cdot) - \widetilde{Q}(s, \cdot, \cdot)\|_\infty = \gamma \cdot \left\| \sum_{s'} p(s' \mid s, \cdot, \cdot) \big(V(s') - \widetilde{V}(s') \big) \right\|_\infty$$
$$= \gamma \cdot \|V - \widetilde{V}\|_\infty$$

Planning: Value Iteration

- Example: Two-player zero-sum SGs (Shapley, 1953)
  - For infinite-horizon case, $\mathcal{B}$ is γ-contracting:

$$\|\mathcal{B}(V) - \mathcal{B}(\widetilde{V})\|_\infty \leq \gamma \cdot \|V - \widetilde{V}\|_\infty$$

  - Key: non-expansiveness of max min operator. For all $s \in \mathcal{S}$

$$\begin{aligned}
&V(s) - \widetilde{V}(s) \\
&= \left| \max_{\mu \in \Delta(\mathcal{A}^1)} \min_{\nu \in \Delta(\mathcal{A}^2)} \mathbb{E}_{a^1 \sim \mu, a^2 \sim \nu} \left[Q(s, a^1, a^2) \right] - \max_{\mu \in \Delta(\mathcal{A}^1)} \min_{\nu \in \Delta(\mathcal{A}^2)} \mathbb{E}_{a^1 \sim \mu, a^2 \sim \nu} \left[\widetilde{Q}(s, a^1, a^2) \right] \right| \\
&\leq \|Q(s, \cdot, \cdot) - \widetilde{Q}(s, \cdot, \cdot)\|_\infty = \gamma \cdot \left\| \sum_{s'} p(s' \mid s, \cdot, \cdot) \big(V(s') - \widetilde{V}(s') \big) \right\|_\infty \\
&= \gamma \cdot \|V - \widetilde{V}\|_\infty
\end{aligned}$$

- Thus, (minimax) value iteration (Shapley, 1953), $V^{k+1} \leftarrow \mathcal{B}(V^k)$, converges to (the unique NE) value V_* as $k \to \infty$
  - NE policy can then be extracted by solving for each $s \in \mathcal{S}$:

$$\big(\pi^1_*(s), \pi^2_*(s)\big) \in \arg \max_{\mu \in \Delta(\mathcal{A}^1)} \min_{\nu \in \Delta(\mathcal{A}^2)} [Q_*(s, \mu, \nu)]$$

Planning: Value Iteration

- Example: n-player general-sum SGs (Fink et al., 1964; Takahashi, 1964)
  - Equilibria are not unique in general, even for matrix-game case

Planning: Value Iteration

- Example: n-player general-sum SGs (Fink et al., 1964; Takahashi, 1964)
  - Equilibria are not unique in general, even for matrix-game case
  - Then, VI proceeds as follows:

$$V_*^{i,h}(s) \leftarrow r^{i,h}(s, \pi_*^h) + \gamma \cdot \sum_{s'} p^h(s' \mid s, \pi_*^h) V_*^{i,h+1}(s')$$

where π_*^h comes from $\mathtt{Equilibrum} \in \{\mathtt{NE}, \mathtt{CE}, \mathtt{CCE}\}$ oracle for matrix games ($\mathtt{NE}$ is $\mathtt{PPAD}$-hard to compute (Daskalakis et al., 2009; Chen et al., 2009); $\mathtt{CE}, \mathtt{CCE}$ are tractable by solving LPs)

$$\pi_*^h \in \mathtt{Equilibrum}\left[\left[r^{i,h}(s, \cdot) + \gamma \cdot \sum_{s'} p^h(s' \mid s, \cdot) V_*^{i,h+1}(s')\right]_{i \in [n]}\right]$$

Planning: Value Iteration

- Example: n-player general-sum SGs (Fink et al., 1964; Takahashi, 1964)
  - Equilibria are not unique in general, even for matrix-game case
  - Then, VI proceeds as follows:

$$V_*^{i,h}(s) \leftarrow r^{i,h}(s, \pi_*^h) + \gamma \cdot \sum_{s'} p^h(s' \mid s, \pi_*^h) V_*^{i,h+1}(s')$$

where π_*^h comes from $\texttt{Equilibrum} \in \{\texttt{NE}, \texttt{CE}, \texttt{CCE}\}$ oracle for matrix games ($\texttt{NE}$ is $\texttt{PPAD}$-hard to compute (Daskalakis et al., 2009; Chen et al., 2009); $\texttt{CE}, \texttt{CCE}$ are tractable by solving LPs)

$$\pi_*^h \in \texttt{Equilibrum}\left[\left[r^{i,h}(s,\cdot) + \gamma \cdot \sum_{s'} p^h(s' \mid s, \cdot) V_*^{i,h+1}(s')\right]_{i\in[n]}\right]$$

  - Can also define VI for Q-function

$$Q_*^{i,h}(s, a^1, \cdots, a^n) \leftarrow r^{i,h}(s, a^1, \cdots, a^n) + \gamma \cdot \sum_{s'} p^h(s' \mid s, a^1, \cdots, a^n) \cdot Q_*^{i,h+1}(s', \pi_*^{h+1})$$

Planning: Value Iteration

- Example: n-player general-sum SGs (Fink et al., 1964; Takahashi, 1964)
  - Equilibria are not unique in general, even for matrix-game case
  - Then, VI proceeds as follows:

$$V_*^{i,h}(s) \leftarrow r^{i,h}(s, \pi_*^h) + \gamma \cdot \sum_{s'} p^h(s' \mid s, \pi_*^h) V_*^{i,h+1}(s')$$

    where π_*^h comes from $\texttt{Equilibrum} \in \{\texttt{NE}, \texttt{CE}, \texttt{CCE}\}$ oracle for matrix games ($\texttt{NE}$ is PPAD-hard to compute (Daskalakis et al., 2009; Chen et al., 2009); $\texttt{CE}, \texttt{CCE}$ are tractable by solving LPs)

$$\pi_*^h \in \texttt{Equilibrum}\left[\left[r^{i,h}(s, \cdot) + \gamma \cdot \sum_{s'} p^h(s' \mid s, \cdot) V_*^{i,h+1}(s')\right]_{i \in [n]}\right]$$

  - Can also define VI for Q-function

$$Q_*^{i,h}(s, a^1, \cdots, a^n) \leftarrow r^{i,h}(s, a^1, \cdots, a^n) + \gamma \cdot \sum_{s'} p^h(s' \mid s, a^1, \cdots, a^n) \cdot Q_*^{i,h+1}(s', \pi_*^{h+1})$$

- Finite-horizon: stops in H steps; infinite-horizon: no γ-contracting in general!
- For infinite-horizon: non-stationary equilibrium is easy to compute; stationary equilibrium may(?) be hard (come back later)

Planning: Policy Iteration

Planning: Policy Iteration

- Finite-horizon: essentially the same as VI (exercise!)

Planning: Policy Iteration

- Finite-horizon: essentially the same as VI (exercise!)
- Infinite-horizon is more subtle, even for two-player zero-sum/minimax case: naive PI (Pollatschek and Avi-Itzhak, 1969) as follows does not converge in general (Van Der Wal, 1978; Condon, 1990)

Policy evaluation: $$V^{k+1}(s) = \mathcal{B}^{\infty}_{\pi^{1,k},\pi^{2,k}}(V^k)(s)$$

where $\mathcal{B}_{\pi^1,\pi^2}(V)(s) := r(s,\pi^1(s),\pi^2(s)) + \gamma \cdot p(\cdot \mid s,\pi^1(s),\pi^2(s)) \cdot V,$

Policy improvement ("Greedy" step):

$$\big(\pi^{1,k+1}(s),\pi^{2,k+1}(s)\big) \in \max_{\mu\in\Delta(\mathcal{A}^1)} \min_{\nu\in\Delta(\mathcal{A}^2)} \big[r(s,\mu,\nu) + \gamma \cdot p(\cdot \mid s,\mu,\nu) \cdot V^{k+1}\big]$$

Planning: Policy Iteration

- Provable convergent variant (Hoffman and Karp, 1966):
  - Computation heavy: solve $\Omega\left(\frac{1}{1-\gamma}\right)$ MDPs (Hansen et al., 2013)

Policy evaluation: $V^{k+1}(s) = \mathcal{B}^{\infty}_{\pi^{1,k}}(V^k)(s)$

where $\mathcal{B}_{\pi^1}(V)(s) := \min_{\nu \in \Delta(\mathcal{A}^2)} \left[r(s, \pi^1(s), \nu) + \gamma \cdot p(\cdot \mid s, \pi^1(s), \nu) \cdot V \right],$

Policy improvement ("Greedy" step):

$$\left(\pi^{1,k+1}(s), \pi^{2,k+1}(s)\right) \in \max_{\mu \in \Delta(\mathcal{A}^1)} \min_{\nu \in \Delta(\mathcal{A}^2)} \left[r(s, \mu, \nu) + \gamma \cdot p(\cdot \mid s, \mu, \nu) \cdot V^{k+1} \right]$$

- Other convergent variants with lighter computation (but maybe higher space complexity) (Filar and Tolwinski, 1991; Bertsekas, 2021; Brahma et al., 2022; Winnicki and Srikant, 2023)

Planning: Policy Iteration

- In general, policy-based algorithms can be hard to converge for games: no value monotonicity (key to single-agent PI convergence) due to agents' conflict objectives
  - Usually need some asymmetric update rules between agents, to obtain monotonicity (Hoffman and Karp, 1966; Condon, 1990; Filar and Tolwinski, 1991; Patek, 1997; Bertsekas, 2021; Brahma et al., 2022)
  - Will see more later in learning settings!

Planning: (Nonlinear) Programming

- In contrast to single-agent MDP, there is no LP in general, but a nonlinear program for characterizing NE:

Planning: (Nonlinear) Programming

- In contrast to single-agent MDP, there is no LP in general, but a nonlinear program for characterizing NE:

$$\min_{\pi, \{v^i\}_{i\in[n]}} \quad \sum_{i\in[n]} \rho^\top \left(v^i - (I - \gamma p(\pi))^{-1} r^i(\pi)\right) \quad \boxed{\text{Nash gap}}$$

$$s.t. \quad v^i(s) \geq r^i(s, a^i, \pi^{-i}) + \gamma p(\cdot \mid s, a^i, \pi^{-i}) \cdot v^i, \quad \forall\, s, a^i, i \qquad \boxed{\text{best-response}}$$

$$\pi^i(s) \in \Delta(\mathcal{A}^i), \quad \forall\, s, i \qquad \boxed{\text{simplex constraints}}$$

Planning: (Nonlinear) Programming

- In contrast to single-agent MDP, there is no LP in general, but a nonlinear program for characterizing NE:

$$\min_{\pi,\{v^i\}_{i\in[n]}} \quad \sum_{i\in[n]} \rho^\top \left(v^i - (I-\gamma p(\pi))^{-1} r^i(\pi)\right) \quad \boxed{\text{Nash gap}}$$

$$s.t. \quad v^i(s) \geq r^i(s, a^i, \pi^{-i}) + \gamma p(\cdot \mid s, a^i, \pi^{-i}) \cdot v^i, \quad \forall\, s, a^i, i \quad \boxed{\text{best-response}}$$

$$\pi^i(s) \in \Delta(\mathcal{A}^i), \quad \forall\, s, i \quad \boxed{\text{simplex constraints}}$$

- Can be made as a LP for single-controller and other special SGs, and a sequence of LPs for turn-based SGs (Filar and Vrieze, 2012)

Part I.B: Classical Results

Learning: Value-based Algorithms

- MARL: finding solutions with data and no (full) model knowledge

Learning: Value-based Algorithms

- MARL: finding solutions with data and no (full) model knowledge
- Most earlier multi-agent RL algorithms are value-based
- Minimax Q-learning (Littman, 1994) for two-player zero-sum SGs:
  - Require solving a min max at each iteration, via e.g., LP

$$
\begin{aligned}
Q^{k+1}(s_k, a_k^1, a_k^2) \leftarrow & (1-\alpha_k) \cdot Q^k(s_k, a_k^1, a_k^2) \\
& + \alpha_k \cdot \left[r(s_k, a_k^1, a_k^2) + \gamma \cdot \max_{\mu \in \Delta(\mathcal{A}^1)} \min_{\nu \in \Delta(\mathcal{A}^2)} \left[Q^k(s_{k+1}, \mu, \nu) \right] \right]
\end{aligned}
$$

Learning: Value-based Algorithms

- MARL: finding solutions with data and no (full) model knowledge
- Most earlier multi-agent RL algorithms are value-based
- Minimax Q-learning (Littman, 1994) for two-player zero-sum SGs:
  - Require solving a min max at each iteration, via e.g., LP

$$
\begin{aligned}
Q^{k+1}(s_k, a_k^1, a_k^2) \leftarrow & (1-\alpha_k) \cdot Q^k(s_k, a_k^1, a_k^2) \\
& + \alpha_k \cdot \Big[r(s_k, a_k^1, a_k^2) + \gamma \cdot \max_{\mu \in \Delta(\mathcal{A}^1)} \min_{\nu \in \Delta(\mathcal{A}^2)} \big[Q^k(s_{k+1}, \mu, \nu) \big] \Big]
\end{aligned}
$$

- A Stochastic Approximation of the corresponding value iteration:

$$
Q_*^h(s, a^1, a^2) \leftarrow r(s, a^1, a^2) + \gamma \cdot p(\cdot \mid s, a^1, a^2) \cdot \max_{\mu \in \Delta(\mathcal{A}^1)} \min_{\nu \in \Delta(\mathcal{A}^2)} \big[Q_*^{h+1}(\cdot, \mu, \nu) \big]
$$

Learning: Value-based Algorithms

- Convergence guarantee:

Theorem (Littman and Szepesvári (1996); Szepesvári and Littman (1999))

Suppose every state s is visited infinitely often during minimax-Q-learning, and stepsizes $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\sum_{k=1}^{\infty} (\alpha_k)^2 < \infty$, then Q^k converges to the NE Q-value $Q_* = Q_{\pi_*}$ as $k \to \infty$.

Learning: Value-based Algorithms

- Convergence guarantee:

Theorem (Littman and Szepesvári (1996); Szepesvári and Littman (1999))

Suppose every state s is visited infinitely often during minimax-Q-learning, and stepsizes $\sum_{k=1}^{\infty} \alpha_k = \infty$ *and* $\sum_{k=1}^{\infty} (\alpha_k)^2 < \infty$*, then* Q^k *converges to the NE Q-value* $Q_* = Q_{\pi_*}$ *as* $k \to \infty$*.*

- Key: γ-contracting of $\mathcal{B}$; similar to single-agent Q-learning (Watkins and Dayan, 1992; Jaakkola et al., 1993; Tsitsiklis, 1994)
- In fact, (Szepesvári and Littman, 1999) provided a unified analysis framework as long as the iterating (Bellman) operator is contracting

Learning: Value-based Algorithms

- Extend to general-sum – Nash Q-learning (Hu and Wellman, 2003):
  - Each agent need to maintain all agents' Q-function estimates
  - Require solving an NE for a general-sum game at each iteration (computationally intractable)
  - Only converge under very restricted assumptions (Bowling, 2000); again, to ensure the contracting property of NE

$$Q^{i,k+1}(s_k, a_k^1, \cdots, a_k^n) \leftarrow (1-\alpha_k) \cdot Q^{i,k}(s_k, a_k^1, \cdots, a_k^n) + \alpha_k \cdot \Big(r^i(s_k, a_k^1, \cdots, a_k^n) + \gamma \cdot \Big[\mathtt{NE} \Big[\big\{ Q^{i,k}(s_{k+1}, \cdot) \big\}_{i\in[n]} \Big] \Big]^i \Big)$$

Learning: Value-based Algorithms

- Extend to general-sum – Nash Q-learning (Hu and Wellman, 2003):
  - Each agent need to maintain all agents' Q-function estimates
  - Require solving an NE for a general-sum game at each iteration (computationally intractable)
  - Only converge under very restricted assumptions (Bowling, 2000); again, to ensure the contracting property of $\mathtt{NE}$

$$Q^{i,k+1}(s_k, a_k^1, \cdots, a_k^n) \leftarrow (1-\alpha_k) \cdot Q^{i,k}(s_k, a_k^1, \cdots, a_k^n) + \alpha_k \cdot \left(r^i(s_k, a_k^1, \cdots, a_k^n) + \gamma \cdot \left[\mathtt{NE}\left[\left\{ Q^{i,k}(s_{k+1}, \cdot) \right\}_{i \in [n]} \right] \right]^i \right)$$

- Friend-or-Foe Q-learning (Littman, 2001): replace $\mathtt{NE}\left[\left\{ Q^{i,k} \right\}_{i \in [n]} \right]$ by

$$\max_{\mu \in \Delta\left(\prod_{j \in \mathtt{Friends}} \mathcal{A}^j\right)} \min_{(a^\ell \in \mathcal{A}^\ell)_{\ell \in \mathtt{Foes}}} Q^{i,k}\left(\cdot, \mu, (a^\ell)_{\ell \in \mathtt{Foes}}\right)$$

  - Always converge; to NE if it is either adversarial or coordination

Learning: Value-based Algorithms

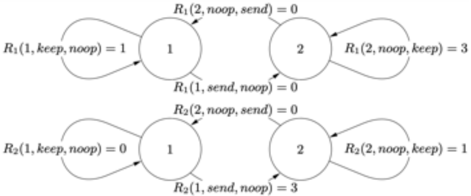
- Other variants: correlated Q-learning (Greenwald et al., 2003) for general-sum SGs; Q-learning (Arslan and Yüksel, 2017) for Teams and weakly-acyclic SGs...

Learning: Value-based Algorithms

- Other variants: correlated Q-learning (Greenwald et al., 2003) for general-sum SGs; Q-learning (Arslan and Yüksel, 2017) for Teams and weakly-acyclic SGs...
- Are value-based algorithms at odds with (inf-horizon) general-sum SGs?

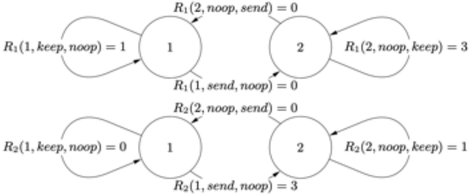
Learning: Value-based Algorithms

- Other variants: correlated Q-learning (Greenwald et al., 2003) for general-sum SGs; Q-learning (Arslan and Yüksel, 2017) for Teams and weakly-acyclic SGs...
- Are value-based algorithms at odds with (inf-horizon) general-sum SGs?
  - (Zinkevich et al., 2006) showed that value iteration (on Q) cannot find stationary equilibrium in arbitrary general-sum SGs
  - Constructed "NoSDE (Nasty) games"



Learning: Value-based Algorithms

- Other variants: correlated Q-learning (Greenwald et al., 2003) for general-sum SGs; Q-learning (Arslan and Yüksel, 2017) for Teams and weakly-acyclic SGs...
- Are value-based algorithms at odds with (inf-horizon) general-sum SGs?
  - (Zinkevich et al., 2006) showed that value iteration (on Q) cannot find stationary equilibrium in arbitrary general-sum SGs
  - Constructed "NoSDE (Nasty) games"

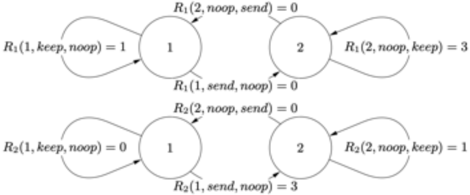


Theorem (Zinkevich et al. (2006))

Every NoSDE game has a unique stationary equilibrium policy. For any NoSDE game Γ with equilibrium policy π, $\exists$ another NoSDE game Γ' with equilibrium policy π', s.t. $Q_\pi^\Gamma = Q_{\pi'}^{\Gamma'}$, but $\pi \neq \pi'$ and $V_\pi^\Gamma \neq V_{\pi'}^{\Gamma'}$.

Learning: Value-based Algorithms

- Other variants: correlated Q-learning (Greenwald et al., 2003) for general-sum SGs; Q-learning (Arslan and Yüksel, 2017) for Teams and weakly-acyclic SGs...
- Are value-based algorithms at odds with (inf-horizon) general-sum SGs?
  - (Zinkevich et al., 2006) showed that value iteration (on Q) cannot find stationary equilibrium in arbitrary general-sum SGs
  - Constructed "NoSDE (Nasty) games"



Theorem (Zinkevich et al. (2006))

Every NoSDE game has a unique stationary equilibrium policy. For any NoSDE game Γ with equilibrium policy π, $\exists$ another NoSDE game Γ' with equilibrium policy π', s.t. $Q_\pi^\Gamma = Q_{\pi'}^{\Gamma'}$, but $\pi \neq \pi'$ and $V_\pi^\Gamma \neq V_{\pi'}^{\Gamma'}$.

  - Advocated a non-stationary equilibrium concept: cyclic equilibria

Learning: Model-based Algorithms

- Model-based: learn models explicitly, and plan in the learned model

Learning: Model-based Algorithms

- Model-based: learn models explicitly, and plan in the learned model
- E^3 for single-controller SGs (Brafman and Tennenholtz, 2000) and R-Max (Brafman and Tennenholtz, 2002) for general zero-sum SGs
  - R-Max balances exploration-exploitation via *optimism in face of uncertainty* (Lattimore and Szepesvári, 2020; Szepesvári, 2022)
  - Key idea: initialize a model with maximal possible reward $R_{\max}$ to encourage exploration, and update during learning
  - Results: convergence with poly sample and computation complexities (can be high)

Learning: Rationality and Convergence

- We have mostly discussed convergence

Learning: Rationality and Convergence

- We have mostly discussed convergence
- (Bowling and Veloso, 2001) argued that a desirable multi-agent learning algorithm should be both convergent and rational:
    - Rationality: the algorithm converges to its opponent's best response if the opponent converges to a stationary policy
    - I.e., the algorithm can exploit weak opponents

Learning: Rationality and Convergence

- We have mostly discussed convergence
- (Bowling and Veloso, 2001) argued that a desirable multi-agent learning algorithm should be both convergent and rational:
  - Rationality: the algorithm converges to its opponent's best response if the opponent converges to a stationary policy
  - I.e., the algorithm can exploit weak opponents
- Minimax (and Nash, Friend-or-Foe) Q-learning are not rational: they converge to equilibrium regardless of what the opponents play

Learning: Rationality and Convergence

- (Bowling and Veloso, 2001) proposed the WoLF (Win-or-Learn-Fast) principle, provably rational and empirically convergent:

$$Q^i(s, a^i) \leftarrow (1-\alpha)Q^i(s, a^i) + \alpha \left(r^i + \gamma \max_{\tilde{a}^i} Q(s', \tilde{a}^i) \right) \quad \boxed{\text{Q-learning}}$$

$$\bar{\pi}^i(s) \leftarrow \bar{\pi}^i(s) + \frac{1}{N(s)} \left(\pi^i(s) - \bar{\pi}^i(s) \right) \quad \boxed{\text{average policy}}$$

$$\pi^i(s, a^i) \leftarrow \pi^i(s, a^i) + \begin{cases} \delta & \text{if } a^i \in \operatorname{argmax} \; Q^i(s, a^i) \\ \frac{-\delta}{|\mathcal{A}^i|-1} & \text{otherwise} \end{cases} \quad \boxed{\text{sampling policy}}$$

with projection of $\pi^i(s)$ on $\Delta(\mathcal{A}^i)$ and δ satisfying WoLF with $\delta_w < \delta_l$

$$\delta = \begin{cases} \delta_w & \text{if } \sum_{a^i} \pi^i(s, a^i) Q^i(s, a^i) > \sum_{a^i} \bar{\pi}^i(s, a^i) Q^i(s, a^i) \quad \boxed{\text{win}} \\ \delta_l & \text{otherwise} \quad \boxed{\text{learn fast}} \end{cases}$$

- In general, decentralized/independent algorithms (as if a single-agent RL algorithm) are more likely to be rational (come back later)

Part II: Modern Results

Modern MARL Theory

- We may call out AlphaGo (Silver et al., 2016) again, as the watershed

Modern MARL Theory

- We may call out AlphaGo (Silver et al., 2016) again, as the watershed
- What's changed?

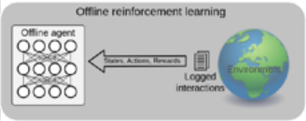
Modern MARL Theory

- We may call out AlphaGo (Silver et al., 2016) again, as the watershed
- What's changed?
  - Non-asymptotic guarantees: regret guarantees, sample complexity, computational complexity
  - Function approximation: inspired by the empirical successes of "deep" (MA)RL
  - New models/settings: beyond canonical stochastic games, with engineering applications

Part II.A: New Guarantees

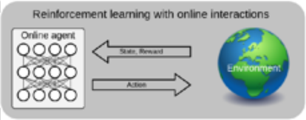
Non-asymptotic Analyses: Sampling Protocols

Simulator



Offline

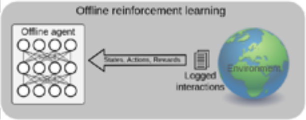
Online



- Simulator setting: good data coverage:
  - Generative model setting (Kearns and Singh, 1999; Kakade, 2003): can sample from any state-action pairs (s, a), e.g., from simulators
  - Trajectory/Markovian sampling with explorative state initialization and/or behavior policies that ensure "all states are visited" (Even-Dar et al., 2003; Beck and Srikant, 2012)

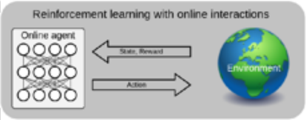
Non-asymptotic Analyses: Sampling Protocols

Simulator



Offline

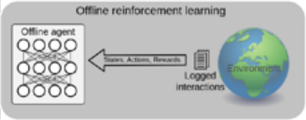
Online



- Simulator setting: good data coverage:
  - Generative model setting (Kearns and Singh, 1999; Kakade, 2003): can sample from any state-action pairs (s, a), e.g., from simulators
  - Trajectory/Markovian sampling with explorative state initialization and/or behavior policies that ensure "all states are visited" (Even-Dar et al., 2003; Beck and Srikant, 2012)
- Online (exploration) setting: no simulator, needs to tradeoff exploration and exploitation through interactions with the environment

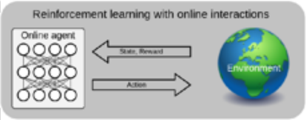
Non-asymptotic Analyses: Sampling Protocols

Simulator



Offline

Online



- Simulator setting: good data coverage:
  - Generative model setting (Kearns and Singh, 1999; Kakade, 2003): can sample from any state-action pairs (s, a), e.g., from simulators
  - Trajectory/Markovian sampling with explorative state initialization and/or behavior policies that ensure "all states are visited" (Even-Dar et al., 2003; Beck and Srikant, 2012)
- Online (exploration) setting: no simulator, needs to tradeoff exploration and exploitation through interactions with the environment
- Offline setting: no interactions allowed, learn from fixed datasets that may not have full/good coverage

Non-asymptotic Analyses: Metrics

- Simulator and offline settings: sample complexity to achieve

$$\texttt{Equilibrium-Gap}(\pi^{out}) \leq \epsilon$$

that scales as $\texttt{poly}\left(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{\epsilon}, H, \log\left(\frac{1}{\delta}\right)\right)$, with $H \sim \frac{1}{1-\gamma}$

Non-asymptotic Analyses: Metrics

- Simulator and offline settings: sample complexity to achieve

$$\texttt{Equilibrium-Gap}(\pi^{out}) \leq \epsilon$$

that scales as $\texttt{poly}\left(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{\epsilon}, H, \log\left(\frac{1}{\delta}\right)\right)$, with $H \sim \frac{1}{1-\gamma}$

- Online setting: regret

$\boxed{\text{single-agent:}}$ $\texttt{Regret}(K) := \sum_{k=1}^{K} \left[V_*^1(s_{1,k}) - V_{\pi^k}^1(s_{1,k})\right]$

$\boxed{\text{two-agent zero-sum:}}$ $\texttt{Regret}(K) := \sum_{k=1}^{K} \left(V_{\dagger,\pi^{2,k}}^1(s_{1,k}) - V_{\pi^{1,k},\dagger}^1\right)(s_{1,k})$

$\boxed{\text{n-agent:}}$ $\texttt{Regret}_{\{\texttt{NE},\texttt{CCE}\}}(K) := \sum_{k=1}^{K} \max_{i\in[n]} \left(V_{\dagger,\pi^{-i,k}}^{i,1}(s_{1,k}) - V_{\pi^k}^{i,1}\right)(s_{1,k})$

depending on π^k is product or correlated; and $\texttt{Regret}_{\texttt{CE}}$ is defined w.r.t. $\max_{\phi^i} V_{(\phi^i\diamond\pi^{i,k})\odot\pi^{-i,k}}^{i,1}(s_{1,k})$

- Goal: achieve $\texttt{Regret}(K) \sim o(K)$ and $\texttt{poly}(|\mathcal{S}|, |\mathcal{A}|, H, \log(1/\delta))$

Non-asymptotic Analyses: Metrics

- Simulator and offline settings: sample complexity to achieve

$$\texttt{Equilibrium-Gap}(\pi^{out}) \leq \epsilon$$

that scales as $\texttt{poly}\left(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{\epsilon}, H, \log\left(\frac{1}{\delta}\right)\right)$, with $H \sim \frac{1}{1-\gamma}$

- Online setting: regret

$\boxed{\text{single-agent:}}$ $\texttt{Regret}(K) := \sum_{k=1}^{K} \left[V_*^1(s_{1,k}) - V_{\pi^k}^1(s_{1,k})\right]$

$\boxed{\text{two-agent zero-sum:}}$ $\texttt{Regret}(K) := \sum_{k=1}^{K} \left(V_{\dagger,\pi^{2,k}}^1(s_{1,k}) - V_{\pi^{1,k},\dagger}^1\right)(s_{1,k})$

$\boxed{\text{n-agent:}}$ $\texttt{Regret}_{\{\texttt{NE},\texttt{CCE}\}}(K) := \sum_{k=1}^{K} \max_{i\in[n]} \left(V_{\dagger,\pi^{-i,k}}^{i,1}(s_{1,k}) - V_{\pi^k}^{i,1}\right)(s_{1,k})$

depending on π^k is product or correlated; and $\texttt{Regret}_{\texttt{CE}}$ is defined w.r.t. $\max_{\phi^i} V^{i,1}_{(\phi^i\diamond\pi^{i,k})\odot\pi^{-i,k}}(s_{1,k})$

- Goal: achieve $\texttt{Regret}(K) \sim o(K)$ and $\texttt{poly}(|\mathcal{S}|, |\mathcal{A}|, H, \log(1/\delta))$
- If $|\mathcal{A}| = \prod_{i\in[n]} |\mathcal{A}^i|$ is replaced by $\max_{i\in[n]} |\mathcal{A}^i|$, it is even polynomial in n, and thus "breaks the curse of multi-agents" (Jin et al., 2023a)

Simulator Setting: Model-based Algorithms

- For any $(s, a^1, \cdots, a^n)$, one can sample $s' \sim p(\cdot \mid s, a^1, \cdots, a^n)$
- Can "plug-in" any black-box planning oracles, e.g., VI, PI, etc.
- Mitigate non-stationarity issue due to all agents' adapting

$$\hat{p}(\cdot \mid s, a^1, a^2) = \frac{N(s,a^1,a^2,s')}{N(s,a^1,a^2)}$$

Simulator Setting: Model-based Algorithms

- For any $(s, a^1, \cdots, a^n)$, one can sample $s' \sim p(\cdot \mid s, a^1, \cdots, a^n)$
- Can "plug-in" any black-box planning oracles, e.g., VI, PI, etc.
- Mitigate non-stationarity issue due to all agents' adapting

$$\hat{p}(\cdot \mid s, a^1, a^2) = \frac{N(s,a^1,a^2,s')}{N(s,a^1,a^2)}$$

Theorem (ZKBY, '20,'23)

This model-based MARL algorithm is near minimax optimal in the generative model setting, with sample complexity $\widetilde{\mathcal{O}}(|S||A^1||A^2|(1-\gamma)^{-3}\epsilon^{-2})$*, and lower bound* $\widetilde{\mathcal{O}}(|S|(|A^1|+|A^2|)(1-\gamma)^{-3}\epsilon^{-2})$*. Moreover, when reward is given after estimating* $\hat{p}$*, both upper and lower bounds are* $\widetilde{\mathcal{O}}(|S||A^1||A^2|(1-\gamma)^{-3}\epsilon^{-2})$ *and model-based MARL is thus minimax optimal in this case.*

Simulator Setting: Model-based Algorithms

- For any $(s, a^1, \cdots, a^n)$, one can sample $s' \sim p(\cdot \mid s, a^1, \cdots, a^n)$
- Can "plug-in" any black-box planning oracles, e.g., VI, PI, etc.
- Mitigate non-stationarity issue due to all agents' adapting

$$\hat{p}(\cdot \mid s, a^1, a^2) = \frac{N(s, a^1, a^2, s')}{N(s, a^1, a^2)}$$

Theorem (ZKBY, '20,'23)
This model-based MARL algorithm is near minimax optimal in the generative model setting, with sample complexity $\widetilde{\mathcal{O}}(|S||A^1||A^2|(1-\gamma)^{-3}\epsilon^{-2})$, and lower bound $\widetilde{\mathcal{O}}(|S|(|A^1|+|A^2|)(1-\gamma)^{-3}\epsilon^{-2})$. Moreover, when reward is given after estimating $\hat{p}$, both upper and lower bounds are $\widetilde{\mathcal{O}}(|S||A^1||A^2|(1-\gamma)^{-3}\epsilon^{-2})$ and model-based MARL is thus minimax optimal in this case.

- Shows the unique separation of model-based approach in MARL
  - Power: generalize to multiple rewards/tasks (after $\hat{p}$ estimated)
  - Limitation: less adaptive and thus suboptimal in $|A^1|, |A^2|$

Simulator Setting: Model-based Algorithms

- For any $(s, a^1, \cdots, a^n)$, one can sample $s' \sim p(\cdot \mid s, a^1, \cdots, a^n)$
- Can "plug-in" any black-box planning oracles, e.g., VI, PI, etc.
- Mitigate non-stationarity issue due to all agents' adapting

$$\hat{p}(\cdot \mid s, a^1, a^2) = \frac{N(s,a^1,a^2,s')}{N(s,a^1,a^2)}$$

Theorem (ZKBY, '20,'23)
This model-based MARL algorithm is near minimax optimal in the generative model setting, with sample complexity $\widetilde{\mathcal{O}}(|S||A^1||A^2|(1-\gamma)^{-3}\epsilon^{-2})$*, and lower bound* $\widetilde{\mathcal{O}}(|S|(|A^1|+|A^2|)(1-\gamma)^{-3}\epsilon^{-2})$*. Moreover, when reward is given after estimating* $\hat{p}$*, both upper and lower bounds are* $\widetilde{\mathcal{O}}(|S||A^1||A^2|(1-\gamma)^{-3}\epsilon^{-2})$ *and model-based MARL is thus minimax optimal in this case.*

- Shows the unique separation of model-based approach in MARL
  - Power: generalize to multiple rewards/tasks (after $\hat{p}$ estimated)
  - Limitation: less adaptive and thus suboptimal in $|A^1|, |A^2|$
- (Subramanian et al., 2023): general-sum $\widetilde{\mathcal{O}}(|S|\prod_{i\in[n]}|A^i|(1-\gamma)^{-3}\epsilon^{-2})$

Simulator Setting: Model-based Algorithms

- For any $(s, a^1, \cdots, a^n)$, one can sample $s' \sim p(\cdot \mid s, a^1, \cdots, a^n)$
- Can "plug-in" any black-box planning oracles, e.g., VI, PI, etc.
- Mitigate non-stationarity issue due to all agents' adapting

$$\hat{p}(\cdot \mid s, a^1, a^2) = \frac{N(s,a^1,a^2,s')}{N(s,a^1,a^2)}$$

Theorem (ZKBY, '20,'23)

This model-based MARL algorithm is near minimax optimal in the generative model setting, with sample complexity $\widetilde{\mathcal{O}}(|S||A^1||A^2|(1-\gamma)^{-3}\epsilon^{-2})$, and lower bound $\widetilde{\mathcal{O}}(|S|(|A^1|+|A^2|)(1-\gamma)^{-3}\epsilon^{-2})$. Moreover, when reward is given after estimating $\hat{p}$, both upper and lower bounds are $\widetilde{\mathcal{O}}(|S||A^1||A^2|(1-\gamma)^{-3}\epsilon^{-2})$ and model-based MARL is thus minimax optimal in this case.

- Shows the unique separation of model-based approach in MARL
  - Power: generalize to multiple rewards/tasks (after $\hat{p}$ estimated)
  - Limitation: less adaptive and thus suboptimal in $|A^1|, |A^2|$
- (Subramanian et al., 2023): general-sum $\widetilde{\mathcal{O}}(|S|\prod_{i\in[n]}|A^i|(1-\gamma)^{-3}\epsilon^{-2})$
- Q: minimax optimality + break curse of multi-agents simultaneously?

Simulator Setting: Value-based Algorithms

- (Sidford et al., 2020): generalize variance-reduced Q-learning to attained minimax-optimal for two-player zero-sum turn-based SGs

$$\widetilde{\mathcal{O}}\left(\frac{|S|\cdot\max_{i=1,2}\{|A^i|\}}{(1-\gamma)^3\epsilon^2}\right)$$

- (Gao et al., 2021): Q-learning of (Arslan and Yüksel, 2017) for weakly-acyclic general-sum SGs
- (Lee, 2023): minimax Q-learning under explorative behavior policies/reachability assumption

Simulator Setting: Value-based Algorithms

- (Sidford et al., 2020): generalize variance-reduced Q-learning to attained minimax-optimal for two-player zero-sum turn-based SGs

$$\widetilde{\mathcal{O}}\left(\frac{|S|\cdot\max_{i=1,2}\{|A^i|\}}{(1-\gamma)^3\epsilon^2}\right)$$

- (Gao et al., 2021): Q-learning of (Arslan and Yüksel, 2017) for weakly-acyclic general-sum SGs
- (Lee, 2023): minimax Q-learning under explorative behavior policies/reachability assumption
- (Li et al., 2022) addressed our open question in previous slide: Q-learning with Follow-the-Regularized-Leader (FTRL) + variance-aware bonus

$$\mathcal{O}\left(\frac{H^4|\mathcal{S}|\sum_{i\in[n]}|\mathcal{A}^i|}{\epsilon^2}\right)$$

for NE/CCE in non-stationary finite SGs

Simulator Setting: Policy-based Algorithms

- (Li et al., 2022)'s FTRL part is kind-of policy-based (inherent connection to natural policy gradient (Agarwal et al., 2021))
- (Winnicki and Srikant, 2023): lookahead policy iteration (to fix naive PI) + [ZKBY, '20, '23] for two-player zero-sum SGs

Online Setting: Model-based Algorithms

- One key idea to tradeoff exploration-exploitation: optimism in face of uncertainty (OFU) principle (Szepesvári, 2022)
- Maintain optimistic estimates of values/models to encourage exploration

Online Setting: Model-based Algorithms

- One key idea to tradeoff exploration-exploitation: optimism in face of uncertainty (OFU) principle (Szepesvári, 2022)
- Maintain optimistic estimates of values/models to encourage exploration
- Model-based algorithms:
  - Optimistic value iteration (Bai and Jin, 2020; Liu et al., 2021):

$$\bar{Q}^{i,h}(s, a^1, \cdots, a^n) \leftarrow \min\left\{(r^{i,h} + \hat{p}^h \bar{V}^{i,h+1})(s, a^1, \cdots, a^n) + \beta_t, H\right\}$$

$$\underline{Q}^{i,h}(s, a^1, \cdots, a^n) \leftarrow \min\left\{(r^{i,h} + \hat{p}^h \underline{V}^{i,h+1})(s, a^1, \cdots, a^n) - \beta_t, 0\right\}$$

$$\pi^h(s) \leftarrow \texttt{Equilibrium}\left(\bar{Q}^{1,h}(s,\cdot), \cdots, \bar{Q}^{n,h}(s,\cdot)\right)$$

with $\bar{V}^{i,h}(s) = \bar{Q}^{i,h}(s, \pi^h(s))$, $\underline{V}^{i,h}(s) = \underline{Q}^{i,h}(s, \pi^h(s))$, and $\hat{p}^h(\cdot \mid s_h, a_h) = N_h(s_h, a_h, \cdot)/N_h(s_h, a_h)$

Online Setting: Model-based Algorithms

- One key idea to tradeoff exploration-exploitation: optimism in face of uncertainty (OFU) principle (Szepesvári, 2022)
- Maintain optimistic estimates of values/models to encourage exploration
- Model-based algorithms:
  - Optimistic value iteration (Bai and Jin, 2020; Liu et al., 2021):

$$\bar{Q}^{i,h}(s,a^1,\cdots,a^n) \leftarrow \min\left\{(r^{i,h}+\hat{p}^h\bar{V}^{i,h+1})(s,a^1,\cdots,a^n)+\beta_t, H\right\}$$
$$\underline{Q}^{i,h}(s,a^1,\cdots,a^n) \leftarrow \min\left\{(r^{i,h}+\hat{p}^h\underline{V}^{i,h+1})(s,a^1,\cdots,a^n)-\beta_t, 0\right\}$$
$$\pi^h(s) \leftarrow \texttt{Equilibrium}\left(\bar{Q}^{1,h}(s,\cdot),\cdots,\bar{Q}^{n,h}(s,\cdot)\right)$$

    with $\bar{V}^{i,h}(s)=\bar{Q}^{i,h}(s,\pi^h(s))$, $\underline{V}^{i,h}(s)=\underline{Q}^{i,h}(s,\pi^h(s))$, and $\hat{p}^h(\cdot\,|\,s_h,a_h)=N_h(s_h,a_h,\cdot)/N_h(s_h,a_h)$

  - Think of zero-sum case — OFU for both min and max players
  - Small differences in bonus-term choices and `Equilibrium oracle` for the zero-sum case: (Bai and Jin, 2020) used NE and (Liu et al., 2021) used CCE (see also (Xie et al., 2020))

Online Setting: Model-based Algorithms

- Guarantee of optimistic VI:

Theorem (Liu et al. (2021))

This optimistic VI algorithm achieves

$$\mathtt{Regret}_{\{NE,CE,CCE\}}(K) \sim \widetilde{\mathcal{O}}\left(\sqrt{H^4|\mathcal{S}|^2 \prod_{i\in[n]} |\mathcal{A}^i| K}\right),$$

and outputs a Markov policy π^{out} *that is an* ϵ-$\{\mathtt{NE,CE,CCE}\}$, *i.e.,*

$$\{\mathtt{NE,CE,CCE}\}\text{-}\mathtt{Gap}(\pi^{out}) \leq \epsilon$$

in $\widetilde{\mathcal{O}}\left(\frac{H^4|\mathcal{S}|^2 \prod_{i\in[n]} |\mathcal{A}^i|}{\epsilon^2}\right)$ *episodes.*

- Better bound of $\widetilde{\mathcal{O}}\left(\frac{H^3|\mathcal{S}||\mathcal{A}^1||\mathcal{A}^2|}{\epsilon^2}\right)$ for two-player zero-sum case with different bonus terms (Liu et al., 2021)

Online Setting: Model-based Algorithms

- Guarantee of optimistic VI:

Theorem (Liu et al. (2021))

This optimistic VI algorithm achieves

$$\mathtt{Regret}_{\{NE,CE,CCE\}}(K) \sim \widetilde{\mathcal{O}}\left(\sqrt{H^4|\mathcal{S}|^2 \prod_{i\in[n]} |\mathcal{A}^i| K}\right),$$

and outputs a Markov policy π^{out} *that is an* ϵ-$\{NE, CE, CCE\}$, *i.e.,*

$$\{NE, CE, CCE\}\text{-}Gap(\pi^{out}) \leq \epsilon$$

in $\widetilde{\mathcal{O}}\left(\frac{H^4|\mathcal{S}|^2 \prod_{i\in[n]} |\mathcal{A}^i|}{\epsilon^2}\right)$ *episodes.*

- Better bound of $\widetilde{\mathcal{O}}\left(\frac{H^3|\mathcal{S}||\mathcal{A}^1||\mathcal{A}^2|}{\epsilon^2}\right)$ for two-player zero-sum case with different bonus terms (Liu et al., 2021)
- Lower bound $\Omega\left(\frac{H^3|\mathcal{S}| \max_{i=1,2} |\mathcal{A}^i|}{\epsilon^2}\right)$; similar gap as in generative model
  - Can "the curse of multi-agents" also be broken in online setting?

Online Setting: Value-based Algorithms

- Optimistic Nash V-learning (Bai et al., 2020; Jin et al., 2023a):

$$\bar{V}^h(s_h) \leftarrow (1-\alpha_t)\bar{V}^h(s_h) + \alpha_t \left(r^h + V^{h+1}(s_{h+1}) + \beta_t\right)$$

$$\pi^h(s_h) \leftarrow \texttt{Adv-Bandit}\left(a_h, \frac{H - r^h - V^{h+1}(s_{h+1})}{H}\right)$$

  with $V^h(s_h) \leftarrow \min\{H+1-h, \bar{V}^h(s_h)\}$ and `Adv-Bandit` an adversarial bandit algorithm, e.g., EXP3 (Lattimore and Szepesvári, 2020)

- First proposed in (Bai et al., 2020) for zero-sum SGs, then generalized to general-sum SGs as "V-learning" (Jin et al., 2023a); see also (Song et al., 2022; Mao and Başar, 2022)

Online Setting: Value-based Algorithms

- Optimistic Nash V-learning (Bai et al., 2020; Jin et al., 2023a):

$$\bar{V}^h(s_h) \leftarrow (1-\alpha_t)\bar{V}^h(s_h) + \alpha_t \left(r^h + V^{h+1}(s_{h+1}) + \beta_t\right)$$

$$\pi^h(s_h) \leftarrow \texttt{Adv-Bandit}\left(a_h, \frac{H - r^h - V^{h+1}(s_{h+1})}{H}\right)$$

  with $V^h(s_h) \leftarrow \min\{H+1-h, \bar{V}^h(s_h)\}$ and `Adv-Bandit` an adversarial bandit algorithm, e.g., EXP3 (Lattimore and Szepesvári, 2020)

- First proposed in (Bai et al., 2020) for zero-sum SGs, then generalized to general-sum SGs as "V-learning" (Jin et al., 2023a); see also (Song et al., 2022; Mao and Başar, 2022)

Theorem (Jin et al. (2023a))

V-learning can output a non-Markov policy π^{out} that is an ϵ-NE/CCE in $\widetilde{\mathcal{O}}\left(\frac{H^5|\mathcal{S}|\max_{i\in[n]}|\mathcal{A}^i|}{\epsilon^2}\right)$ episodes. A monotonic variant can output a Markov policy that is an ϵ-NE for two-player zero-sum SGs with the same sample complexity.

- Replacing `Adv-Bandit` oracle by a no-swap-regret one can address CE
- V-learning breaks "the curse" in finite-horizon online setting

Online Setting

- Other notable results:
  - (Wang et al., 2023) and [CZD, '23]: break "the curse" with independent linear function approximation
  - (Wei et al., 2017): a model-based one for average-reward SGs, based on UCRL2 (Jaksch et al., 2010)
  - (Xie et al., 2020; Chen et al., 2022c): linear function approximation for the game model
  - (Jin et al., 2022; Huang et al., 2022; Xiong et al., 2022; Foster et al., 2023a; Liu et al., 2024): general function approximation

Offline Setting

- Dataset: $\mathcal{D} := \left\{ (s_h^{(\ell)}, a_h^{(\ell)}, r^{h,(\ell)}, s_{h+1}^{(\ell)}) \right\}_{\ell \in [N], h \in [H]} \sim d_\mu$
- When offline data has full coverage, batch RL on the dataset works (pay distribution shift coefficient) (Munos and Szepesvári, 2008; Chen and Jiang, 2019)
- Interesting regime: partial data coverage

Offline Setting

- Dataset: $\mathcal{D} := \left\{(s_h^{(\ell)}, a_h^{(\ell)}, r^{h,(\ell)}, s_{h+1}^{(\ell)})\right\}_{\ell\in[N],h\in[H]} \sim d_\mu$
- When offline data has full coverage, batch RL on the dataset works (pay distribution shift coefficient) (Munos and Szepesvári, 2008; Chen and Jiang, 2019)
- Interesting regime: partial data coverage
- What is the minimal the offline data distribution d_μ should cover?
  - For single-agent RL, single optimal policy π_* coverage suffices (Jin et al., 2021; Rashidinejad et al., 2021; Xie et al., 2021b; Zhan et al., 2022), [OPZZ, '22]

$$\max_{s,a} \frac{d_\rho^{\pi_*}(s,a)}{d_\mu(s,a)} \leq C < \infty$$

Offline Setting

- For multi-agent RL, Nash equilibrium coverage is not enough; unilateral coverage is required (Cui and Du, 2022b)

$$\checkmark \ \max_{s,a} \frac{d_\rho^{\pi_*^1,\pi_*^2}(s,a)}{d_\mu(s,a)} \le C, \quad \text{✗} \ \max\left\{ \max_{s,a,\pi^2} \frac{d_\rho^{\pi_*^1,\pi^2}(s,a)}{d_\mu(s,a)}, \max_{s,a,\pi^1} \frac{d_\rho^{\pi^1,\pi_*^2}(s,a)}{d_\mu(s,a)} \right\} \le C$$

Min Player

Max Player

	b_1	b_2	...	b_B
a_1	[0.4, 0.6]	[0.8, 1]	...	[0.7, 0.8]
a_2	[0, 0.1]	[0.4, 0.7]	...	[0.6, 0.7]
...	...	...	...	...
a_A	[0.1, 0.3]	[0.2, 0.4]	...	...

Offline Setting

- For multi-agent RL, Nash equilibrium coverage is not enough; unilateral coverage is required (Cui and Du, 2022b)

$$\checkmark \ \max_{s,a} \frac{d_\rho^{\pi_*^1,\pi_*^2}(s,a)}{d_\mu(s,a)} \le C, \quad \text{✘} \ \max\left\{ \max_{s,a,\pi^2} \frac{d_\rho^{\pi_*^1,\pi^2}(s,a)}{d_\mu(s,a)}, \max_{s,a,\pi^1} \frac{d_\rho^{\pi^1,\pi_*^2}(s,a)}{d_\mu(s,a)} \right\} \le C$$

Min Player

Max Player

	b_1	b_2	...	b_B
a_1	[0.4, 0.6]	[0.8, 1]	...	[0.7, 0.8]
a_2	[0, 0.1]	[0.4, 0.7]	...	[0.6, 0.7]
...	...	...	...	...
a_A	[0.1, 0.3]	[0.2, 0.4]	...	...

- Under unilateral coverage, pessimistic Nash value iteration is efficient (Cui and Du, 2022b,a); see also (Zhong et al., 2022)

Part II.B: New Models

Beyond Canonical SGs: Multi-player Zero-sum SGs

- For matrix games: computationally, for NE, general-sum is hard (Daskalakis et al., 2009; Chen et al., 2009); two-player zero-sum is easy
- Is there a class of games in between that is also easy (in some sense)?

Beyond Canonical SGs: Multi-player Zero-sum SGs

- For matrix games: computationally, for NE, general-sum is hard (Daskalakis et al., 2009; Chen et al., 2009); two-player zero-sum is easy
- Is there a class of games in between that is also easy (in some sense)?
- Multi-player zero-sum games:
  - Naively, 3-player zero-sum is hard (with a dummy player)
  - With a polymatrix payoff structure (Cai et al., 2016) (below for agent i and some graph $\mathcal{G} := ([n], \mathcal{E})$), it enjoys equilibrium collapse: CCE=NE

$$r^i(a) = \sum_{j:(i,j)\in\mathcal{E}} r^{i,j}(a^i, a^j)$$

Beyond Canonical SGs: Multi-player Zero-sum SGs

- For matrix games: computationally, for NE, general-sum is hard (Daskalakis et al., 2009; Chen et al., 2009); two-player zero-sum is easy
- Is there a class of games in between that is also easy (in some sense)?
- Multi-player zero-sum games:
  - Naively, 3-player zero-sum is hard (with a dummy player)
  - With a polymatrix payoff structure (Cai et al., 2016) (below for agent i and some graph $\mathcal{G} := ([n], \mathcal{E})$), it enjoys equilibrium collapse: CCE=NE

$$r^i(a) = \sum_{j:(i,j)\in\mathcal{E}} r^{i,j}(a^i, a^j)$$

- What about stochastic games?

Beyond Canonical SGs: Multi-player Zero-sum SGs

- For matrix games: computationally, for NE, general-sum is hard (Daskalakis et al., 2009; Chen et al., 2009); two-player zero-sum is easy
- Is there a class of games in between that is also easy (in some sense)?
- Multi-player zero-sum games:
  - Naively, 3-player zero-sum is hard (with a dummy player)
  - With a polymatrix payoff structure (Cai et al., 2016) (below for agent i and some graph $\mathcal{G} := ([n], \mathcal{E})$), it enjoys equilibrium collapse: CCE=NE

$$r^i(a) = \sum_{j:(i,j)\in\mathcal{E}} r^{i,j}(a^i, a^j)$$

- What about stochastic games?
- One can define this polymatrix structure for each auxiliary game's payoff induced by any value vector V [P*Z*O, '23]:

$$Q_V^i(s,a) := r^i(s,a) + \gamma \sum_{s'} p(s' \mid s, a) V(s') = \sum_{j:(i,j)\in\mathcal{E}} Q_V^{i,j}(a^i, a^j)$$

  - It covers polymatrix reward + single-controller/turn-based/additive structures (Flesch et al., 2007)

Beyond Canonical SGs: Multi-player Zero-sum SGs

- Markov CCE collapses to Markov NE [P*Z*O, '23]
  - Non-stationary NE can be easy (by finding non-stationary CCE)

Beyond Canonical SGs: Multi-player Zero-sum SGs

- Markov CCE collapses to Markov NE [P*Z*O, '23]
  - Non-stationary NE can be easy (by finding non-stationary CCE)
- Concurrent work (Kalogiannis and Panageas, 2023) defines a different model: polymatrix reward + switching controller transition
  - Different techniques for equilibrium collapse, based on the nonlinear program introduced in Part I

Beyond Canonical SGs: Stochastic/Markov Potential Games

- We have mostly talked about "non-cooperative" settings, what about "(near-)cooperative" ones?

Beyond Canonical SGs: Stochastic/Markov Potential Games

- We have mostly talked about "non-cooperative" settings, what about "(near-)cooperative" ones?
- Some early definitions in (Marden, 2012; Macua et al., 2018); recently, Markov potential games (Leonardos et al., 2022; Zhang et al., 2024a): there exists a potential function Φ s.t. for each state s and all agents i

$$\Phi_{\pi^i,\pi^{-i}}(s) - \Phi_{\widetilde{\pi}^i,\pi^{-i}}(s) = V^i_{\pi^i,\pi^{-i}}(s) - V^i_{\widetilde{\pi}^i,\pi^{-i}}(s)$$

- Potential reward $\not\Leftrightarrow$ Markov potential game (Leonardos et al., 2022)
  - This model thus addresses mixed cooperative/competitive agents

Beyond Canonical SGs: Linear Quadratic Dynamic Games

- The "tabular case" for continuous space settings
- Two-player zero-sum linear quadratic (LQ) dynamic games:

$$r(s,a,b) = -s^\top Q s - a^\top R^1 a + b^\top R^2 b,$$
$$s_{h+1} = A s_h + B^1 a_h + B^2 b_h + w_h$$

Beyond Canonical SGs: Linear Quadratic Dynamic Games

- The "tabular case" for continuous space settings
- Two-player zero-sum linear quadratic (LQ) dynamic games:

$$r(s, a, b) = -s^\top Q s - a^\top R^1 a + b^\top R^2 b,$$
$$s_{h+1} = A s_h + B^1 a_h + B^2 b_h + w_h$$

- An old model (Başar and Bernhard, 1995), receives increasing attention in MARL in recent years [ZYB, '19; ZHB, '20] and (Bu et al., 2019; Wu et al., 2023)
- Has a deep connection to risk-sensitive control and $\mathcal{H}_\infty$ robust control (Whittle, 1981; Başar and Bernhard, 1995)

Beyond Canonical SGs: Linear Quadratic Dynamic Games

- The "tabular case" for continuous space settings
- Two-player zero-sum linear quadratic (LQ) dynamic games:

$$r(s,a,b) = -s^\top Q s - a^\top R^1 a + b^\top R^2 b,$$
$$s_{h+1} = A s_h + B^1 a_h + B^2 b_h + w_h$$

- An old model (Başar and Bernhard, 1995), receives increasing attention in MARL in recent years [ZYB, '19; ZHB, '20] and (Bu et al., 2019; Wu et al., 2023)
- Has a deep connection to risk-sensitive control and $\mathcal{H}_\infty$ robust control (Whittle, 1981; Başar and Bernhard, 1995)
- General-sum case can also be defined (Başar and Olsder, 1998; Mazumdar et al., 2020; Hambly et al., 2023; Aggarwal et al., 2024), as well as the potential case (Hosseinirad et al., 2024)

Beyond SGs: Networked/Distributed MARL

- Non-game-theoretic cooperative setting: a group of networked agents

$$\max_{\{\pi^i\}_{i\in[n]}} \quad \mathbb{E}\left[\sum_{t\geq 0}\gamma^t\left(\frac{1}{n}\sum_{i\in[n]} r_t^i\right)\right]$$

with neighbor-to-neighbor communications




Beyond SGs: Networked/Distributed MARL

- Non-game-theoretic cooperative setting: a group of networked agents

$$\max_{\{\pi^i\}_{i\in[n]}} \quad \mathbb{E}\left[\sum_{t\geq 0}\gamma^t\left(\frac{1}{n}\sum_{i\in[n]} r_t^i\right)\right]$$

  with neighbor-to-neighbor communications

- Centralized ✔ v.s. Distributed/Networked ✘

  - Scalable to large-number of agents
  - Resilient to attacks
  - Better preserve the privacy of each agent
  - Distributed/Consensus optimization for static problems (Xiao et al., 2007; Nedic and Ozdaglar, 2009; Duchi et al., 2011)

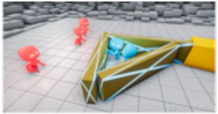
Beyond SGs: Networked/Distributed MARL

- For dynamic decision-making problems:
  - (Kar et al., 2013) for Q-learning; [ZYLZB, '18] for actor-critic
  - Many followups (Wai et al., 2018; Doan et al., 2019, 2021; Lee et al., 2018; Chu et al., 2019; Figura et al., 2021; Zhang and Zavlanos, 2019; Sun et al., 2020; Stanković et al., 2023)
- Recent advances: (Qu et al., 2020, 2022; Zhang et al., 2023; Zhou et al., 2023; Olsson et al., 2024)
  - With additional locality assumptions on the reward/transition $\Longrightarrow$ local policies suffice

Beyond SGs: Other Models

- Partially-observable SGs:
  - In practice, the system state is almost never observable
  - Additionally, each agent may not have other agents' observations – asymmetric information structure/decentralized decision-making

$$o_t^i \sim \mathcal{O}^i(\cdot \mid s_t), \qquad a_t^i \sim \pi^{i,t}(\cdot \mid o_1^i, a_1^i, o_2^i, a_2^i, \cdots, o_t^i)$$

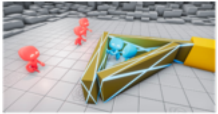


  - Many known (computational) hardness results (Witsenhausen, 1968; Tsitsiklis and Athans, 1985) from the Control literature

Beyond SGs: Other Models

- Partially-observable SGs:
  - In practice, the system state is almost never observable
  - Additionally, each agent may not have other agents' observations – asymmetric information structure/decentralized decision-making

$$o_t^i \sim \mathcal{O}^i(\cdot \mid s_t), \qquad a_t^i \sim \pi^{i,t}(\cdot \mid o_1^i, a_1^i, o_2^i, a_2^i, \cdots, o_t^i)$$



  - Many known (computational) hardness results (Witsenhausen, 1968; Tsitsiklis and Athans, 1985) from the Control literature
  - Recently, (Liu et al., 2022a; Qiu et al., 2024) focused on sample-efficiency (polynomial sample complexities)
  - Further, [LZ, '23] established (quasi-)polynomial sample and computation complexities, by exploiting the "information-sharing" formalism from decentralized stochastic control (Mahajan, 2008; Nayyar et al., 2013b,a)

Beyond SGs: Other Models

- Team setting: one-vs-team (adversarial team Markov games) (Kalogiannis et al., 2023)

  - Efficient computation algorithm for ϵ-stationary Nash equilibrium

Beyond SGs: Other Models

- Mean-field setting: large population of agents with interactions through mean-field state/population distribution $\mu \in \Delta(\mathcal{S})$

$$r^i(s,a) \Longrightarrow r(s,a,\mu), \qquad p(s' \mid s,a) \Longrightarrow p(s' \mid s,a,\mu)$$

  - Provable mean-field RL (Guo et al., 2019; Perrin et al., 2020; Xie et al., 2021a; Cui and Koeppl, 2021; Pérolat et al., 2022; Geist et al., 2022; Anahtarci et al., 2023; Guo et al., 2023a; Yardim et al., 2023; Huang et al., 2024b,a; Ramponi et al., 2024)
  - Computation: it can be PPAD-hard with only Lipschitz dynamics and rewards (Yardim et al., 2024)

Part II.C: New Algorithm Class: Policy Optimization for MARL

New Algorithm Class: Policy Optimization

- In practice, policy gradient/optimization methods, e.g., proximal policy optimization (PPO) (Schulman et al., 2017), are very useful (default)
- Recent advances in understanding policy gradient (PG) methods (Cai et al., 2020; Wang et al., 2020; Agarwal et al., 2021; Bhandari and Russo, 2024; Cen et al., 2022; Fatkhullin et al., 2023) and many more

New Algorithm Class: Policy Optimization

- In practice, policy gradient/optimization methods, e.g., proximal policy optimization (PPO) (Schulman et al., 2017), are very useful (default)
- Recent advances in understanding policy gradient (PG) methods (Cai et al., 2020; Wang et al., 2020; Agarwal et al., 2021; Bhandari and Russo, 2024; Cen et al., 2022; Fatkhullin et al., 2023) and many more
- Policy gradient methods for MARL: parameterize each agent's policy π^i as $\pi^i_{\theta^i}$, and run gradient ascent

$$\theta^i_{k+1} \leftarrow \theta^i_k + \alpha_k \cdot \nabla_{\theta^i} V^i\left(\theta^i_k, \theta^{-i}_k\right)$$

where $V^i\left(\theta^i_k, \theta^{-i}_k\right) := \mathbb{E}_{s_0 \sim \rho} V^i_{\pi^i_{\theta^i}, \pi^{-i}_{\theta^{-i}}}(s_0)$ and the PG $\nabla_{\theta^i} V^i\left(\theta^i_k, \theta^{-i}_k\right)$ can be estimated by samples

New Algorithm Class: Policy Optimization

- In practice, policy gradient/optimization methods, e.g., proximal policy optimization (PPO) (Schulman et al., 2017), are very useful (default)
- Recent advances in understanding policy gradient (PG) methods (Cai et al., 2020; Wang et al., 2020; Agarwal et al., 2021; Bhandari and Russo, 2024; Cen et al., 2022; Fatkhullin et al., 2023) and many more
- Policy gradient methods for MARL: parameterize each agent's policy π^i as $\pi^i_{\theta^i}$, and run gradient ascent

$$\theta^i_{k+1} \leftarrow \theta^i_k + \alpha_k \cdot \nabla_{\theta^i} V^i \left(\theta^i_k, \theta^{-i}_k\right)$$

where $V^i \left(\theta^i_k, \theta^{-i}_k\right) := \mathbb{E}_{s_0 \sim \rho} V^i_{\pi^i_{\theta^i}, \pi^{-i}_{\theta^{-i}}}(s_0)$ and the PG $\nabla_{\theta^i} V^i \left(\theta^i_k, \theta^{-i}_k\right)$ can be estimated by samples

- Policy gradient theorem (Sutton et al., 2000) for SGs:

$$\nabla_{\theta^i} V^i \left(\theta^i, \theta^{-i}\right) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\theta, a^i \sim \pi^i_{\theta^i}(\cdot \mid s)} \left[\nabla_{\theta^i} \log \pi^i_{\theta^i}(a^i \mid s) \cdot q^i_{\pi_\theta}(s, a^i)\right]$$

New Algorithm Class: Policy Optimization

- In practice, policy gradient/optimization methods, e.g., proximal policy optimization (PPO) (Schulman et al., 2017), are very useful (default)
- Recent advances in understanding policy gradient (PG) methods (Cai et al., 2020; Wang et al., 2020; Agarwal et al., 2021; Bhandari and Russo, 2024; Cen et al., 2022; Fatkhullin et al., 2023) and many more
- Policy gradient methods for MARL: parameterize each agent's policy π^i as $\pi^i_{\theta^i}$, and run gradient ascent
$$\theta^i_{k+1} \leftarrow \theta^i_k + \alpha_k \cdot \nabla_{\theta^i} V^i\left(\theta^i_k, \theta^{-i}_k\right)$$
where $V^i\left(\theta^i_k, \theta^{-i}_k\right) := \mathbb{E}_{s_0 \sim \rho} V^i_{\pi^i_{\theta^i}, \pi^{-i}_{\theta^{-i}}}(s_0)$ and the PG $\nabla_{\theta^i} V^i\left(\theta^i_k, \theta^{-i}_k\right)$ can be estimated by samples
- Policy gradient theorem (Sutton et al., 2000) for SGs:
$$\nabla_{\theta^i} V^i\left(\theta^i, \theta^{-i}\right) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\theta, a^i \sim \pi^i_{\theta^i}(\cdot \mid s)} \left[\nabla_{\theta^i} \log \pi^i_{\theta^i}(a^i \mid s) \cdot q^i_{\pi_\theta}(s, a^i)\right]$$
which, under direct parameterization $\theta^i_s = \pi^i_{\theta^i}(s) \in \Delta(\mathcal{A}^i)$, reduces to
$$\nabla_{\theta^i_{s,a^i}} V^i\left(\theta^i, \theta^{-i}\right) = \frac{1}{1-\gamma} d_\theta(s) q^i_{\pi_\theta}(s, a^i)$$

Partial Gradient Dominance Property

- A simple but useful fact — "partial" gradient-dominance: assume $d_\theta(\cdot) > 0$ (simulator setting; good data coverage; it holds if $\rho(\cdot) > 0$)

Partial Gradient Dominance Property

- A simple but useful fact — "partial" gradient-dominance: assume $d_\theta(\cdot) > 0$ (simulator setting; good data coverage; it holds if $\rho(\cdot) > 0$)

$$V^i\left(\underbrace{\widetilde{\theta}^i, \theta^{-i}}_{\widetilde{\theta}}\right) - V^i\left(\theta^i, \theta^{-i}\right) \overset{\boxed{\text{performance difference lemma (Kakade and Langford, 2002)}}}{=} \frac{1}{1-\gamma}\sum_{s,a} d_{\widetilde{\theta}}(s)\pi_{\widetilde{\theta}}(a \mid s)\left[Q^i_{\pi_\theta}(s,a) - V^i_{\pi_\theta}(s)\right]$$

$$= \frac{1}{1-\gamma}\sum_{s,a^i} d_{\widetilde{\theta}}(s)\pi^i_{\widetilde{\theta}^i}(a^i \mid s)\left[q^i_{\pi_\theta}(s,a^i) - V^i_{\pi_\theta}(s)\right]$$

$$\leq \frac{1}{1-\gamma}\left\|\frac{d_{\widetilde{\theta}}}{d_\theta}\right\|_\infty \sum_s d_\theta(s)\max_{a^i}\left[q^i_{\pi_\theta}(s,a^i) - V^i_{\pi_\theta}(s)\right]$$

$$= \frac{1}{1-\gamma}\left\|\frac{d_{\widetilde{\theta}}}{d_\theta}\right\|_\infty \max_{\bar{\theta}^i \in \Delta(\mathcal{A}^i)^{|\mathcal{S}|}} \sum_s d_\theta(s)\left[q^i_{\pi_\theta}(s,\pi^i_{\bar{\theta}^i}(s)) - V^i_{\pi_\theta}(s)\right]$$

$$= \left\|\frac{d_{\widetilde{\theta}}}{d_\theta}\right\|_\infty \max_{\bar{\theta}^i \in \Delta(\mathcal{A}^i)^{|\mathcal{S}|}} \sum_{s,a^i}\left[(\pi^i_{\bar{\theta}^i} - \pi^i_{\theta^i})(a^i \mid s)\cdot q^i_{\pi_\theta}(s,a^i)\frac{d_\theta(s)}{1-\gamma}\right]$$

$$= \left\|\frac{d_{\widetilde{\theta}}}{d_\theta}\right\|_\infty \max_{\bar{\theta}^i \in \Delta(\mathcal{A}^i)^{|\mathcal{S}|}} (\bar{\theta}^i - \theta^i)\cdot \nabla_{\theta^i} V^i(\theta) \quad \boxed{\leq 0 \text{ is 1st-order opt. cond. fixing } \theta^{-i}}$$

see also [ZYB '19], (Mazumdar et al., 2019) (for LQ cases) and (Daskalakis et al., 2020; Leonardos et al., 2022; Zhang et al., 2024a)

Policy Optimization for Two-player Zero-sum SGs

- The former result implies:

$$\boxed{\text{1st-order stationary point } \theta_* \implies \theta_*^i \text{ best-responds to } \theta_*^{-i},\ \forall\ i \implies \text{NE } \theta_*}$$

Policy Optimization for Two-player Zero-sum SGs

- The former result implies:

$$\boxed{\text{1st-order stationary point } \theta_* \implies \theta_*^i \text{ best-responds to } \theta_*^{-i},\ \forall\, i \implies \text{NE } \theta_*}$$

- Not easy for zero-sum, as $\max_{\theta^1} \min_{\theta^2} V(\theta^1, \theta^2)$ is nonconvex-nonconcave

Policy Optimization for Two-player Zero-sum SGs

- The former result implies:

$$\boxed{\text{1st-order stationary point } \theta_* \implies \theta_*^i \text{ best-responds to } \theta_*^{-i}, \ \forall\, i \implies \text{NE } \theta_*}$$

- Not easy for zero-sum, as $\max_{\theta^1}\min_{\theta^2} V(\theta^1, \theta^2)$ is nonconvex-nonconcave
- (Daskalakis et al., 2020): policy gradient for two-player zero-sum SGs

$$\theta_{k+1}^1 \leftarrow \texttt{Proj}\left[\theta_k^1 + \alpha \cdot \nabla_{\theta^1} V\left(\theta_k^1, \theta_k^2\right)\right]$$
$$\theta_{k+1}^2 \leftarrow \texttt{Proj}\left[\theta_k^2 - \beta \cdot \nabla_{\theta^2} V\left(\theta_k^1, \theta_k^2\right)\right]$$

with $\alpha \asymp \epsilon^{10.5}$ and $\beta \asymp \epsilon^6$, i.e., $\alpha \ll \beta$

  - Asymmetric stepsizes between the two players
  - With asymmetric (player 1) convergence to ϵ-NE in $\texttt{poly}$ samples

Policy Optimization for Two-player Zero-sum SGs

- The former result implies:

$$\boxed{\text{1st-order stationary point } \theta_* \implies \theta_*^i \text{ best-responds to } \theta_*^{-i},\ \forall\, i \implies \text{NE } \theta_*}$$

- Not easy for zero-sum, as $\max_{\theta^1} \min_{\theta^2} V(\theta^1, \theta^2)$ is nonconvex-nonconcave
- (Daskalakis et al., 2020): policy gradient for two-player zero-sum SGs

$$\theta_{k+1}^1 \leftarrow \texttt{Proj}\left[\theta_k^1 + \alpha \cdot \nabla_{\theta^1} V\left(\theta_k^1, \theta_k^2\right)\right]$$
$$\theta_{k+1}^2 \leftarrow \texttt{Proj}\left[\theta_k^2 - \beta \cdot \nabla_{\theta^2} V\left(\theta_k^1, \theta_k^2\right)\right]$$

with $\alpha \asymp \epsilon^{10.5}$ and $\beta \asymp \epsilon^6$, i.e., $\alpha \ll \beta$

  - Asymmetric stepsizes between the two players
  - With asymmetric (player 1) convergence to ϵ-NE in `poly` samples
  - Echoing back to the asymmetry in PI (Hoffman and Karp, 1966; Condon, 1990; Filar and Tolwinski, 1991; Patek, 1997; Bertsekas, 2021; Brahma et al., 2022) (for monotonicity)!

Policy Optimization for Two-player Zero-sum SGs

- Other policy optimization methods that are also asymmetric: (Guo et al., 2021; Zhao et al., 2022; Alacaoglu et al., 2022; Zeng et al., 2022)

Policy Optimization for Two-player Zero-sum SGs

- Other policy optimization methods that are also asymmetric: (Guo et al., 2021; Zhao et al., 2022; Alacaoglu et al., 2022; Zeng et al., 2022)
- Is it possible to have a symmetric one?

Policy Optimization for Two-player Zero-sum SGs

- Other policy optimization methods that are also asymmetric: (Guo et al., 2021; Zhao et al., 2022; Alacaoglu et al., 2022; Zeng et al., 2022)
- Is it possible to have a symmetric one?
- A variant of policy optimization (Wei et al., 2021): optimistic gradient descent-ascent (full-information version)

$$\hat{\pi}^1_{k+1}(s) \leftarrow \texttt{Proj}\left[\hat{\pi}^1_k(s) + \eta Q_k(s, \pi^2_k(s))\right] \quad \boxed{\text{optimistic actor}}$$

$$\pi^1_{k+1}(s) \leftarrow \texttt{Proj}\left[\hat{\pi}^1_{k+1}(s) + \eta Q_k(s, \pi^2_k(s))\right]$$

$$V_k(s) \leftarrow (1-\alpha_k)V_{k-1}(s) + \alpha_k Q_k(s, \pi^1_k(s), \pi^2_k(s)) \quad \boxed{\text{(centralized) smooth critic}}$$

with $Q_k(s, a^1, a^2) := r(s, a^1, a^2) + \gamma \mathbb{E}_{s' \sim p(\cdot \mid, s, a^1, a^2)} V_{k-1}(s')$

Policy Optimization for Two-player Zero-sum SGs

- Other policy optimization methods that are also asymmetric: (Guo et al., 2021; Zhao et al., 2022; Alacaoglu et al., 2022; Zeng et al., 2022)
- Is it possible to have a symmetric one?
- A variant of policy optimization (Wei et al., 2021): optimistic gradient descent-ascent (full-information version)

$$\hat{\pi}^1_{k+1}(s) \leftarrow \texttt{Proj}\left[\hat{\pi}^1_k(s) + \eta Q_k(s, \pi^2_k(s))\right] \qquad \boxed{\text{optimistic actor}}$$

$$\pi^1_{k+1}(s) \leftarrow \texttt{Proj}\left[\hat{\pi}^1_{k+1}(s) + \eta Q_k(s, \pi^2_k(s))\right]$$

$$V_k(s) \leftarrow (1-\alpha_k) V_{k-1}(s) + \alpha_k Q_k(s, \pi^1_k(s), \pi^2_k(s)) \qquad \boxed{\text{(centralized) smooth critic}}$$

with $Q_k(s, a^1, a^2) := r(s, a^1, a^2) + \gamma \mathbb{E}_{s' \sim p(\cdot \,|\, , s, a^1, a^2)} V_{k-1}(s')$

- (Wei et al., 2021): last-iterate convergence rate, symmetric, and rational

Policy Optimization for Two-player Zero-sum SGs

- Other policy optimization methods that are also asymmetric: (Guo et al., 2021; Zhao et al., 2022; Alacaoglu et al., 2022; Zeng et al., 2022)
- Is it possible to have a symmetric one?
- A variant of policy optimization (Wei et al., 2021): optimistic gradient descent-ascent (full-information version)

$$\hat{\pi}^1_{k+1}(s) \leftarrow \texttt{Proj}\left[\hat{\pi}^1_k(s) + \eta Q_k(s, \pi^2_k(s))\right] \qquad \boxed{\text{optimistic actor}}$$

$$\pi^1_{k+1}(s) \leftarrow \texttt{Proj}\left[\hat{\pi}^1_{k+1}(s) + \eta Q_k(s, \pi^2_k(s))\right]$$

$$V_k(s) \leftarrow (1-\alpha_k) V_{k-1}(s) + \alpha_k Q_k(s, \pi^1_k(s), \pi^2_k(s)) \qquad \boxed{\text{(centralized) smooth critic}}$$

with $Q_k(s, a^1, a^2) := r(s, a^1, a^2) + \gamma \mathbb{E}_{s' \sim p(\cdot \mid , s, a^1, a^2)} V_{k-1}(s')$

- (Wei et al., 2021): last-iterate convergence rate, symmetric, and rational
- Other policy optimization methods that are also symmetric, with such a smooth critic framework: (Chen et al., 2022b; Zhang et al., 2022a; Cen et al., 2023; Song et al., 2023; Yang and Ma, 2023; Cai et al., 2024b)
  - Variants on the actor step yield various convergence guarantees: faster rate, last-iterate, Markov sampling, etc.

Policy Optimization for Markov Potential Games

- In contrast, the partial gradient dominance property might be a blessing for the potential case
  - Policy gradient $\Longrightarrow$ gradient descent for (a smooth) potential value function $\Longrightarrow$ conv. to stationary-point $\Longrightarrow$ conv. to NE

Policy Optimization for Markov Potential Games

- In contrast, the partial gradient dominance property might be a blessing for the potential case
  - Policy gradient $\Longrightarrow$ gradient descent for (a smooth) potential value function $\Longrightarrow$ conv. to stationary-point $\Longrightarrow$ conv. to NE
- Indeed, (Leonardos et al., 2022; Zhang et al., 2024a) leveraged this

Policy Optimization for Markov Potential Games

- In contrast, the partial gradient dominance property might be a blessing for the potential case
  - Policy gradient $\Longrightarrow$ gradient descent for (a smooth) potential value function $\Longrightarrow$ conv. to stationary-point $\Longrightarrow$ conv. to NE
- Indeed, (Leonardos et al., 2022; Zhang et al., 2024a) leveraged this
- Other results:
  - Work [DWZJ, '22] took a different route and developed a new second-order performance difference lemma to sharpen the rates and incorporate function approximation
  - Generalization to other policy optimization methods, e.g., natural PG (Fox et al., 2022; Sun et al., 2023) and/or with regularization (Zhang et al., 2022b; Sun et al., 2024), other parameterization (Zhang et al., 2022b), online exploration (Song et al., 2022), average-reward (Cheng et al., 2024), networked (Aydın and Eksin, 2023), and near-potential settings (Guo et al., 2023b)

Policy Optimization for General-sum SGs

- The smooth critic framework can also be generalized to finite-horizon general-sum SGs: (Zhang et al., 2022a; Erez et al., 2023; Cai et al., 2024a; Mao et al., 2024) for (C)CE computation/learning

Policy Optimization for General-sum SGs

- The smooth critic framework can also be generalized to finite-horizon general-sum SGs: (Zhang et al., 2022a; Erez et al., 2023; Cai et al., 2024a; Mao et al., 2024) for (C)CE computation/learning
- Other notable results (both exploit the gradient dominance property):
  - (Anagnostides et al., 2024): ϵ-NE can be efficiently found for single-controller + equilibrium collapse (e.g., two-player or polymatrix zero-sum) cases
  - (Giannou et al., 2022): second-order stationary NE are locally attracting for policy gradient

Policy Optimization for Linear Quadratic Games

- Parameterization w.l.o.g: $a_h = -Ks_h$ and $b_h = -Ls_h$ (Başar and Bernhard, 1995)

Policy Optimization for Linear Quadratic Games

- Parameterization w.l.o.g: $a_h = -Ks_h$ and $b_h = -Ls_h$ (Başar and Bernhard, 1995)

Theorem (ZYB, '19; ZHB, '20; ZZHB, '20)

For two-player zero-sum LQ games, $\max_K \min_L V(K, L)$ is a nonconvex-nonconcave minimax optimization in (K, L), but has the partial gradient dominance property. Also, double-loop policy optimization converges globally to the Nash equilibrium with sublinear rates.

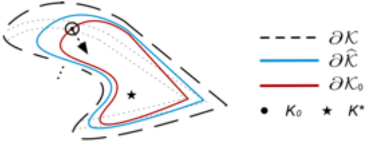
- Double-loop policy optimization:

Algorithm 2 Double-Loop Update

```
for k = 0, ..., K − 1 do
    for l = 0, ..., L − 1 do
        Update L_{l+1} ← PolicyOptimizer(K_k, L_l)
    end for
    Update K_{k+1} ← PolicyOptimizer(K_k, L_L)
end for
```

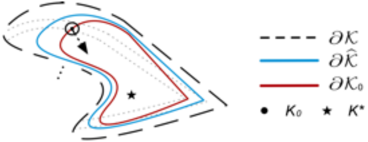
Policy Optimization for Linear Quadratic Games

- Challenges: continuous and unbounded spaces; no global smoothness
- One has to control the iteration path carefully to stay in certain sets



Policy Optimization for Linear Quadratic Games

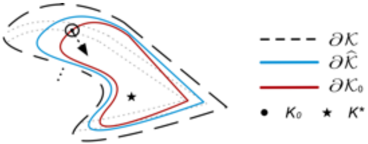
- Challenges: continuous and unbounded spaces; no global smoothness
- One has to control the iteration path carefully to stay in certain sets



- Much faster than existing $\mathcal{H}_\infty$-robust control methods even for computation purposes [ZHB, '20]
  - No need for linear matrix inequality/semi-definite program (Boyd et al., 1994), but just policy parameter-space search – dimension-friendly

Policy Optimization for Linear Quadratic Games

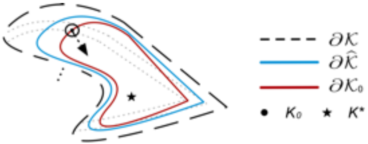
- Challenges: continuous and unbounded spaces; no global smoothness
- One has to control the iteration path carefully to stay in certain sets



- Much faster than existing $\mathcal{H}_\infty$-robust control methods even for computation purposes [ZHB, '20]
  - No need for linear matrix inequality/semi-definite program (Boyd et al., 1994), but just policy parameter-space search – dimension-friendly
- Finite-sample analysis with zeroth-order sampling [ZZHB, '20]
- Recent improved sample complexity in (Wu et al., 2023)

Policy Optimization for Linear Quadratic Games

- Challenges: continuous and unbounded spaces; no global smoothness
- One has to control the iteration path carefully to stay in certain sets

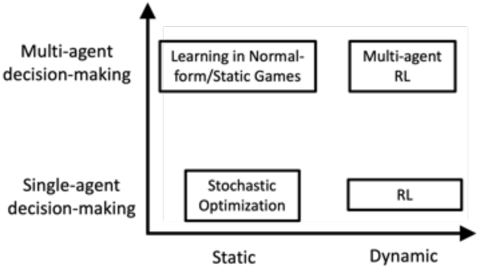


- Much faster than existing $\mathcal{H}_\infty$-robust control methods even for computation purposes [ZHB, '20]
  - No need for linear matrix inequality/semi-definite program (Boyd et al., 1994), but just policy parameter-space search – dimension-friendly
- Finite-sample analysis with zeroth-order sampling [ZZHB, '20]
- Recent improved sample complexity in (Wu et al., 2023)
- Generalization to general-sum settings:
  - Negative (local) convergence result (Mazumdar et al., 2020)
  - Recent advances (Hambly et al., 2023; Aggarwal et al., 2024; Hosseinirad et al., 2024)

Part III: Why Multi-agent RL?
A Learning-in-Games Perspective

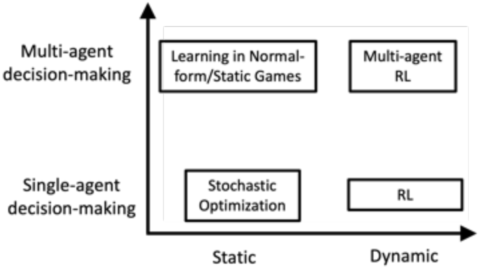
Multi-agent Reinforcement Learning

- Received broad research interest from ML, Econ, Control, and Alg. Game Theory (with an increasing number of workshops/programs at Simons Institute, NeurIPS, ICML, ICLR, CDC ... over the years)
- All these recent exciting advances introduced so far; I personally have contributed to it during Ph.D.



Multi-agent Reinforcement Learning

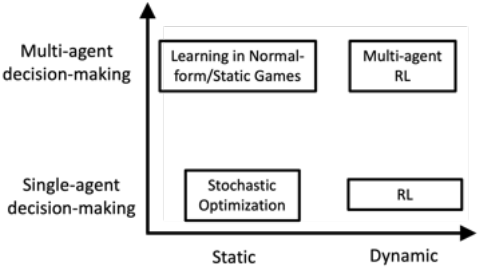
- Received broad research interest from ML, Econ, Control, and Alg. Game Theory (with an increasing number of workshops/programs at Simons Institute, NeurIPS, ICML, ICLR, CDC ... over the years)
- All these recent exciting advances introduced so far; I personally have contributed to it during Ph.D.



- But ...

Multi-agent Reinforcement Learning

- Received broad research interest from ML, Econ, Control, and Alg. Game Theory (with an increasing number of workshops/programs at Simons Institute, NeurIPS, ICML, ICLR, CDC ... over the years)
- All these recent exciting advances introduced so far; I personally have contributed to it during Ph.D.



- But ...

Why multi-agent reinforcement learning?

An Intriguing Question I Had Been Thinking About

If multi-agent learning is the answer, what is the question?

— Yoav Shoham, 2005

An Intriguing Question I Had Been Thinking About

If multi-agent learning is the answer, what is the question?

— Yoav Shoham, 2005

- Polynomial sample & space complexity?
- Online exploration/Offline learning & sublinear regret?
- Faster "equilibrium" computation? Its computational complexity?

An Intriguing Question I Had Been Thinking About

> If multi-agent learning is the answer, what is the question?
>
> — Yoav Shoham, 2005

- Polynomial sample & space complexity?
- Online exploration/Offline learning & sublinear regret?
- Faster "equilibrium" computation? Its computational complexity?

> Does this multi-agent perspective really present new and unique challenges for (sequential) decision-making?

Is “Finding Equilibrium” all We Need/Have Got?

Is "Finding Equilibrium" all We Need/Have Got?

- One traditional explanation of (Nash) equilibrium:

> It results from analysis and introspection by the players, knowing the rules of the game, the rationality of the players, and payoff functions

Is “Finding Equilibrium” all We Need/Have Got?

- One traditional explanation of (Nash) equilibrium:

> It results from analysis and introspection by the players, knowing the rules of the game, the rationality of the players, and payoff functions

- An alternative from Learning-in-Games and Economics literature [Fudenberg & Levine, ’98]:

Is "Finding Equilibrium" all We Need/Have Got?

- One traditional explanation of (Nash) equilibrium:

> It results from analysis and introspection by the players, knowing the rules of the game, the rationality of the players, and payoff functions

- An alternative from Learning-in-Games and Economics literature [Fudenberg & Levine, '98]:

> Equilibrium arises (naturally) as the long-run outcome of a process in which less than fully rational players grope for optimality over time

  - I.e., equilibrium is not the target, but the natural outcome of myopic and non-equilibrating learning dynamics (from each other)

Is "Finding Equilibrium" all We Need/Have Got?

- One traditional explanation of (Nash) equilibrium:

It results from analysis and introspection by the players, knowing the rules of the game, the rationality of the players, and payoff functions

- An alternative from Learning-in-Games and Economics literature [Fudenberg & Levine, '98]:

Equilibrium arises (naturally) as the long-run outcome of a process in which less than fully rational players grope for optimality over time

  - I.e., equilibrium is not the target, but the natural outcome of myopic and non-equilibrating learning dynamics (from each other)
  - The agents may not even realize they are in a game
  - With laboratory evidence (with human participants) – e.g., Nagel's beauty contest experiment [Nagel '95][Duffy and Nagel, '97]
  - "As a 'predictive model' for decision-makers' long-term behaviors"

Is "Finding Equilibrium" all We Need/Have Got?

- One traditional explanation of (Nash) equilibrium:

> It results from analysis and introspection by the players, knowing the rules of the game, the rationality of the players, and payoff functions

- An alternative from Learning-in-Games and Economics literature [Fudenberg & Levine, '98]:

> Equilibrium arises (naturally) as the long-run outcome of a process in which less than fully rational players grope for optimality over time

  - I.e., equilibrium is not the target, but the natural outcome of myopic and non-equilibrating learning dynamics (from each other)
  - The agents may not even realize they are in a game
  - With laboratory evidence (with human participants) – e.g., Nagel's beauty contest experiment [Nagel '95][Duffy and Nagel, '97]
  - "As a 'predictive model' for decision-makers' long-term behaviors"
  - "Learning dynamics is not a computational algorithm"
    - Though as an algorithm, it can be bad/slow!

An Example: Fictitious-Play & Nash Equilibrium

- This perspective has been well-established in normal-form/matrix games, see e.g., [Fudenberg & Levine, '98]

An Example: Fictitious-Play & Nash Equilibrium

- This perspective has been well-established in normal-form/matrix games, see e.g., [Fudenberg & Levine, '98]
- Fictitious-play [Brown, '51]:

Belief Update: For agent i maintains belief $\hat{\pi}_k^{-i}$ at time k,

$$\hat{\pi}_{k+1}^{-i} = \hat{\pi}_k^{-i} + \frac{1}{k} \cdot (a_k^{-i} - \hat{\pi}_k^{-i}),$$

Action Selection: The action a_k^i is taken from best-response

$$a_k^i \in \arg\max_{a^i} \left\{ (a^i)^T Q^i \hat{\pi}_k^{-i} \right\}.$$

An Example: Fictitious-Play & Nash Equilibrium

- This perspective has been well-established in normal-form/matrix games, see e.g., [Fudenberg & Levine, '98]
- Fictitious-play [Brown, '51]:

  Belief Update: For agent i maintains belief $\hat{\pi}_k^{-i}$ at time k,

  $$\hat{\pi}_{k+1}^{-i} = \hat{\pi}_k^{-i} + \frac{1}{k} \cdot (a_k^{-i} - \hat{\pi}_k^{-i}),$$

  Action Selection: The action a_k^i is taken from best-response

  $$a_k^i \in \arg\max_{a^i} \left\{ (a^i)^T Q^i \hat{\pi}_k^{-i} \right\}.$$

  - Nash equilibrium emerges in the long-run, for several classes of matrix games, zero-sum [Robinson, '51], identical-interest [Monderer and Shapley, '96], 2×2 non-zero-sum [Miyasawa, '61]
  - Natural, symmetric, and independent (coordination-free) dynamics

An Example: Fictitious-Play & Nash Equilibrium

- This perspective has been well-established in normal-form/matrix games, see e.g., [Fudenberg & Levine, '98]
- Fictitious-play [Brown, '51]:

  Belief Update: For agent i maintains belief $\hat{\pi}_k^{-i}$ at time k,

  $$\hat{\pi}_{k+1}^{-i} = \hat{\pi}_k^{-i} + \frac{1}{k} \cdot (a_k^{-i} - \hat{\pi}_k^{-i}),$$

  Action Selection: The action a_k^i is taken from best-response

  $$a_k^i \in \arg\max_{a^i} \left\{ (a^i)^T Q^i \hat{\pi}_k^{-i} \right\}.$$

  - Nash equilibrium emerges in the long-run, for several classes of matrix games, zero-sum [Robinson, '51], identical-interest [Monderer and Shapley, '96], 2×2 non-zero-sum [Miyasawa, '61]
  - Natural, symmetric, and independent (coordination-free) dynamics
  - Though it can be slow as a "computational algorithm" [Robinson, '51][Daskalakis and Pan, '14]

A "Learning-in-Games" Perspective of MARL

Is this also true in dynamic games with states/as in RL?

- "Long-run outcome" [Fudenberg & Levine, '98] suggests us to focus on games without reset, i.e., infinite-horizon SGs [Shapley, '53][Fink, '64]

A "Learning-in-Games" Perspective of MARL

Is this also true in dynamic games with states/as in RL?

- "Long-run outcome" [Fudenberg & Levine, '98] suggests us to focus on games without reset, i.e., infinite-horizon SGs [Shapley, '53][Fink, '64]

If "not" in general, maybe it's fine to just embrace it (as a solution concept)?

"In praise of game dynamics" "Let the dynamics show you the way"

— Christo Papadimitriou (at Simons Institute), 2022

On the Other Hand, in (Empirical) Multi-agent RL..

On the Other Hand, in (Empirical) Multi-agent RL..

- Independent learning (IL): each agent runs (variants) of single-agent RL algorithms, to myopically improve her policies, sometimes even oblivious to other agents or the type of the game

On the Other Hand, in (Empirical) Multi-agent RL..

- Independent learning (IL): each agent runs (variants) of single-agent RL algorithms, to myopically improve her policies, sometimes even oblivious to other agents or the type of the game
- Technically, IL is known to suffer from convergence issues [Condon, '90], [Tan, '93], [Claus and Boutilier, '98]
  - Due to the key issue in multi-agent RL: non-stationarity
  - Other agents' learning and adapting processes break the stationary MDP assumption from a single-agent's perspective

On the Other Hand, in (Empirical) Multi-agent RL..

- Independent learning (IL): each agent runs (variants) of single-agent RL algorithms, to myopically improve her policies, sometimes even oblivious to other agents or the type of the game
- Technically, IL is known to suffer from convergence issues [Condon, '90], [Tan, '93], [Claus and Boutilier, '98]
  - Due to the key issue in multi-agent RL: non-stationarity
  - Other agents' learning and adapting processes break the stationary MDP assumption from a single-agent's perspective
- Practically, IL seems to perform well (better than I expected)..

On the Other Hand, in (Empirical) Multi-agent RL..

- Independent learning (IL): each agent runs (variants) of single-agent RL algorithms, to myopically improve her policies, sometimes even oblivious to other agents or the type of the game
- Technically, IL is known to suffer from convergence issues [Condon, '90], [Tan, '93], [Claus and Boutilier, '98]
  - Due to the key issue in multi-agent RL: non-stationarity
  - Other agents' learning and adapting processes break the stationary MDP assumption from a single-agent's perspective
- Practically, IL seems to perform well (better than I expected)..
  - "Is independent learning all you need in the StarCraft multi-agent challenge?" [Witt et al., '20]
  - "The surprising effectiveness of PPO in cooperative multi-agent games," [Yu et al., '21]
  - "Independent algorithms can perform on par with multi-agent ones in cooperative and competitive settings," [Lee et al., '21]
  - "Decentralized reinforcement learning control of a robotic manipulator," [Bușoniu et el., '06]
  - ...

Question of Interest

Question of Interest

Can (Nash) equilibrium be realized by natural and independent learning dynamics in stochastic games?

Question of Interest

Can (Nash) equilibrium be realized by natural and independent learning dynamics in stochastic games?

- If so (in some cases), then it might in turn justify the success of independent learning in multi-agent RL (in certain cases)
- If not (in general), is there any possible fundamental reason?

Independent Learning Made Simple

We identify simple independent learning dynamics that have Nash equilibrium emerge in the long run for certain stochastic games

Independent Learning Made Simple

We identify simple independent learning dynamics that have Nash equilibrium emerge in the long run for certain stochastic games

- The learning dynamics requires no explicit coordination among agents, is symmetric and natural (simple variant of single-agent dynamics, e.g., vanilla independent Q-learning [Claus and Boutilier, '98])
- Each agent is unaware of the type of the game (e.g., zero-sum or not), and sometimes even unaware of the existence of other agents

Decentralized Q-learning Dynamics

- Goal: as similar as vanilla independent Q-learning [Watkins, 89], with no awareness of the opponents' action (set) nor even their existence

Decentralized Q-learning Dynamics

- Goal: as similar as vanilla independent Q-learning [Watkins, 89], with no awareness of the opponents' action (set) nor even their existence
- Recall the local Q function of player i

$$q_\pi^i(s, a^i) := \mathbb{E}_{a^{-i} \sim \pi_s^{-i}} \left\{ Q_\pi^i(s, a^i, a^{-i}) \right\}, \quad \forall \ (s, a^i)$$

Decentralized Q-learning Dynamics

- Goal: as similar as vanilla independent Q-learning [Watkins, 89], with no awareness of the opponents' action (set) nor even their existence
- Recall the local Q function of player i

$$q^i_\pi(s, a^i) := \mathbb{E}_{a^{-i} \sim \pi_s^{-i}} \left\{ Q^i_\pi(s, a^i, a^{-i}) \right\}, \quad \forall \ (s, a^i)$$

- Step 1: Player i infers the opponent's strategy by estimating the local Q-function $q^i_{\pi_k}(s, a^i)$

Decentralized Q-learning Dynamics

- Goal: as similar as vanilla independent Q-learning [Watkins, 89], with no awareness of the opponents' action (set) nor even their existence
- Recall the local Q function of player i

$$q_\pi^i(s, a^i) := \mathbb{E}_{a^{-i} \sim \pi_s^{-i}} \left\{ Q_\pi^i(s, a^i, a^{-i}) \right\}, \quad \forall \; (s, a^i)$$

- Step 1: Player i infers the opponent's strategy by estimating the local Q-function $q_{\pi_k}^i(s, a^i)$
  - Q-learning-type update

$$\hat{q}_{s_k,k+1}^i[a_k^i] = \hat{q}_{s_k,k}^i[a_k^i] + \alpha_{\sharp s_k} \left(r_k^i + \gamma \cdot \hat{v}_{s_{k+1},k}^i - \hat{q}_{s_k,k}^i[a_k^i] \right),$$

Decentralized Q-learning Dynamics

- Goal: as similar as vanilla independent Q-learning [Watkins, 89], with no awareness of the opponents' action (set) nor even their existence
- Recall the local Q function of player i

$$q_\pi^i(s, a^i) := \mathbb{E}_{a^{-i} \sim \pi_s^{-i}} \left\{ Q_\pi^i(s, a^i, a^{-i}) \right\}, \quad \forall \ (s, a^i)$$

- Step 1: Player i infers the opponent's strategy by estimating the local Q-function $q_{\pi_k}^i(s, a^i)$
  - Q-learning-type update

$$\hat{q}_{s_k,k+1}^i[a_k^i] = \hat{q}_{s_k,k}^i[a_k^i] + \alpha_{\sharp s_k} \left(r_k^i + \gamma \cdot \hat{v}_{s_{k+1},k}^i - \hat{q}_{s_k,k}^i[a_k^i] \right),$$

where $a_k^i \sim \bar{\pi}_k^i$, and $\bar{\pi}_k^i$ the smooth best-response w.r.t. $\hat{q}_{s_k,k}^i$:

$$\bar{\pi}_k^i := \underset{\mu \in \Delta(A_{s_k}^i)}{\operatorname{argmax}} \left\{ \mu^T \hat{q}_{s_k,k}^i + \tau_{\sharp s_k} \cdot \nu_{s_k}^i(\mu) \right\}$$

with some perturbation function $\nu_{s_k}^i(\mu)$, e.g., entropy function, and the temperature parameter $\tau_{\sharp s_k} > 0$

Decentralized Q-learning Dynamics

- Goal: as similar as vanilla independent Q-learning [Watkins, 89], with no awareness of the opponents' action (set) nor even their existence
- Recall the local Q function of player i

$$q^i_\pi(s,a^i) := \mathbb{E}_{a^{-i}\sim\pi_s^{-i}}\left\{Q^i_\pi(s,a^i,a^{-i})\right\}, \quad \forall \ (s,a^i)$$

- Step 1: Player i infers the opponent's strategy by estimating the local Q-function $q^i_{\pi_k}(s,a^i)$
  - Q-learning-type update

$$\hat{q}^i_{s_k,k+1}[a^i_k] = \hat{q}^i_{s_k,k}[a^i_k] + \alpha_{\sharp s_k}\left(r^i_k + \gamma\cdot \hat{v}^i_{s_{k+1},k} - \hat{q}^i_{s_k,k}[a^i_k]\right),$$

    where $a^i_k \sim \bar{\pi}^i_k$, and $\bar{\pi}^i_k$ the smooth best-response w.r.t. $\hat{q}^i_{s_k,k}$:

$$\bar{\pi}^i_k := \operatorname*{argmax}_{\mu\in\Delta(A^i_{s_k})}\left\{\mu^T\hat{q}^i_{s_k,k} + \tau_{\sharp s_k}\cdot\nu^i_{s_k}(\mu)\right\}$$

    with some perturbation function $\nu^i_{s_k}(\mu)$, e.g., entropy function, and the temperature parameter $\tau_{\sharp s_k} > 0$
  - Recall: Vanilla independent Q-learning (single-agent dynamics)

$$\hat{q}^i_{s_k,k+1}[a^i_k] = \hat{q}^i_{s_k,k}[a^i_k] + \alpha_{\sharp s_k}\left(r^i_k + \gamma\cdot \max_{a'}\hat{q}^i_{s_{k+1},k}[a'] - \hat{q}^i_{s_k,k}[a^i_k]\right)$$

Decentralized Q-learning Dynamics

- Step 2: Player i estimates the value function $\hat{v}^i_{s,k}$

$$\hat{v}^i_{s_k,k+1} = \hat{v}^i_{s_k,k} + \beta_{\sharp s_k}\left((\bar{\pi}^i_k)^T \hat{q}^i_{s_k,k} - \hat{v}^i_{s_k,k}\right)$$

Decentralized Q-learning Dynamics

- Step 2: Player i estimates the value function $\hat{v}^i_{s,k}$

$$\hat{v}^i_{s_k,k+1} = \hat{v}^i_{s_k,k} + \beta_{\sharp s_k}\left((\bar{\pi}^i_k)^T \hat{q}^i_{s_k,k} - \hat{v}^i_{s_k,k}\right)$$

- All the quantities are maintained locally, without coordination or communication, and symmetric among agents (different from many existing provable MARL algorithms (at that time :)))

Features of the Learning Dynamics

$$\begin{aligned}
\hat{q}^i_{s_k,k+1}[a^i_k] &= \hat{q}^i_{s_k,k}[a^i_k] + \alpha_{\sharp s_k}\left(r^i_k + \gamma \hat{v}^i_{s_{k+1},k} - \hat{q}^i_{s_k,k}[a^i_k]\right) \\
\hat{v}^i_{s_k,k+1} &= \hat{v}^i_{s_k,k} + \beta_{\sharp s_k}\left((\bar{\pi}^i_k)^T \hat{q}^i_{s_k,k} - \hat{v}^i_{s_k,k}\right)
\end{aligned}$$

Features of the Learning Dynamics

$$\hat{q}^i_{s_k,k+1}[a^i_k] = \hat{q}^i_{s_k,k}[a^i_k] + \alpha_{\sharp s_k}\left(r^i_k + \gamma \hat{v}^i_{s_{k+1},k} - \hat{q}^i_{s_k,k}[a^i_k]\right)$$
$$\hat{v}^i_{s_k,k+1} = \hat{v}^i_{s_k,k} + \beta_{\sharp s_k}\left((\bar{\pi}^i_k)^T \hat{q}^i_{s_k,k} - \hat{v}^i_{s_k,k}\right)$$

- Two-timescale: $\lim_{c\to\infty} \frac{\beta_c}{\alpha_c} = 0$, so that the payoffs of the auxiliary game is relatively stationary
  - As if solving an auxiliary normal-form game with payoff matrix

$$\left[r^i_s(a) + \gamma \sum_{s'} p(s' \mid s, a)\hat{v}^i_{s',k}\right]_{a\in A}$$

Features of the Learning Dynamics

$$\hat{q}^i_{s_k,k+1}[a^i_k] = \hat{q}^i_{s_k,k}[a^i_k] + \alpha_{\sharp s_k}\left(r^i_k + \gamma \hat{v}^i_{s_{k+1},k} - \hat{q}^i_{s_k,k}[a^i_k]\right)$$
$$\hat{v}^i_{s_k,k+1} = \hat{v}^i_{s_k,k} + \beta_{\sharp s_k}\left((\bar{\pi}^i_k)^T \hat{q}^i_{s_k,k} - \hat{v}^i_{s_k,k}\right)$$

- Two-timescale: $\lim_{c\to\infty} \frac{\beta_c}{\alpha_c} = 0$, so that the payoffs of the auxiliary game is relatively stationary
  - As if solving an auxiliary normal-form game with payoff matrix

$$\left[r^i_s(a) + \gamma \sum_{s'} p(s' \mid s, a)\hat{v}^i_{s',k}\right]_{a\in A}$$

  - The relatively frozen $\hat{v}^i_{s',k}$ is similar to target network in (deep, single-agent) Q-learning [Mnih et al., '15]

Features of the Learning Dynamics

$$\hat{q}^i_{s_k,k+1}[a^i_k] = \hat{q}^i_{s_k,k}[a^i_k] + \alpha_{\sharp s_k}\left(r^i_k + \gamma \hat{v}^i_{s_{k+1},k} - \hat{q}^i_{s_k,k}[a^i_k]\right)$$
$$\hat{v}^i_{s_k,k+1} = \hat{v}^i_{s_k,k} + \beta_{\sharp s_k}\left((\bar{\pi}^i_k)^T \hat{q}^i_{s_k,k} - \hat{v}^i_{s_k,k}\right)$$

- Two-timescale: $\lim_{c\to\infty} \frac{\beta_c}{\alpha_c} = 0$, so that the payoffs of the auxiliary game is relatively stationary
  - As if solving an auxiliary normal-form game with payoff matrix
    $$\left[r^i_s(a) + \gamma \sum_{s'} p(s' \mid s, a)\hat{v}^i_{s',k}\right]_{a\in A}$$
  - The relatively frozen $\hat{v}^i_{s',k}$ is similar to target network in (deep, single-agent) Q-learning [Mnih et al., '15]
  - Then update $\hat{v}^i_{s',k}$ as the stochastic approximation of minimax value iteration [Shapley, '53] (thus γ-contracting): key to the convergence!

Features of the Learning Dynamics

- This timescale separation may also find evidence in the literature on Evolutionary Game Theory and Behavioral Economics [Ely and Yilankaya '01], [Sandholm '01]: players' choices are more dynamic than their preferences
  - The payoffs in auxiliary games (determined by $\hat{v}^i_{s,k}$) can be viewed as slowly evolving player preferences

Features of the Learning Dynamics

- This timescale separation may also find evidence in the literature on Evolutionary Game Theory and Behavioral Economics [Ely and Yilankaya '01], [Sandholm '01]: players' choices are more dynamic than their preferences
  - The payoffs in auxiliary games (determined by $\hat{v}^i_{s,k}$) can be viewed as slowly evolving player preferences
- Oblivious to the presence of the opponent: radically uncoupled dynamics [Foster & Young, '06]

Convergence Guarantees: Zero-sum Stochastic Games

Theorem (S*Z*LBO, '21)

Under standard assumptions on the stepsizes $\{\alpha_c,\ \beta_c\}_{c\geq 1}$*, certain decreasing rate of the temperature parameter* $\{\tau_c\}_{c\geq 1}$*, and certain reachability assumption of the states, we have*

$$\lim_{k\to\infty} |\hat{v}^i_{s,k} - V^i_{\pi_*}(s)| = 0$$

almost surely. Moreover, the (weighted-)time-average policy of $\{\bar{\pi}^i_k\}_{k\geq 1}$ *also converges to the Nash equilibrium policy almost surely.*

Convergence Guarantees: Zero-sum Stochastic Games

Theorem (S*Z*LBO, '21)

Under standard assumptions on the stepsizes $\{\alpha_c,\ \beta_c\}_{c\geq 1}$*, certain decreasing rate of the temperature parameter* $\{\tau_c\}_{c\geq 1}$*, and certain reachability assumption of the states, we have*

$$\lim_{k\to\infty} |\hat{v}^i_{s,k} - V^i_{\pi_*}(s)| = 0$$

almost surely. Moreover, the (weighted-)time-average policy of $\{\bar{\pi}^i_k\}_{k\geq 1}$ *also converges to the Nash equilibrium policy almost surely.*

- A Corollary: The learning dynamics is (not only convergent but) also rational [Bowling and Veloso '01]

Convergence Guarantees: Zero-sum Stochastic Games

Theorem (S*Z*LBO, '21)

Under standard assumptions on the stepsizes $\{\alpha_c,\ \beta_c\}_{c\geq 1}$*, certain decreasing rate of the temperature parameter* $\{\tau_c\}_{c\geq 1}$*, and certain reachability assumption of the states, we have*

$$\lim_{k\to\infty} |\hat{v}^i_{s,k} - V^i_{\pi_*}(s)| = 0$$

almost surely. Moreover, the (weighted-)time-average policy of $\{\bar{\pi}^i_k\}_{k\geq 1}$ *also converges to the Nash equilibrium policy almost surely.*

- A Corollary: The learning dynamics is (not only convergent but) also rational [Bowling and Veloso '01]
  - "Can exploit a weaker opponent"
  - Thus natural and rational to follow the dynamics in the first place

Convergence Guarantees: Zero-sum Stochastic Games

Theorem (S*Z*LBO, '21)

Under standard assumptions on the stepsizes $\{\alpha_c,\ \beta_c\}_{c\geq 1}$, *certain decreasing rate of the temperature parameter* $\{\tau_c\}_{c\geq 1}$, *and certain reachability assumption of the states, we have*

$$\lim_{k\to\infty} |\hat{v}^i_{s,k} - V^i_{\pi_*}(s)| = 0$$

almost surely. Moreover, the (weighted-)time-average policy of $\{\bar{\pi}^i_k\}_{k\geq 1}$ *also converges to the Nash equilibrium policy almost surely.*

- A Corollary: The learning dynamics is (not only convergent but) also rational [Bowling and Veloso '01]
  - "Can exploit a weaker opponent"
  - Thus natural and rational to follow the dynamics in the first place
- Some finite sample analyses for the double-loop (instead of two-timescale) versions: [CZMOW, '23; '24] and (Ouhamma and Kamgarpour, 2023)

How further can we go with such learning dynamics?

Fictitious-play Property and Game-agnostic Convergence

Fictitious-play Property and Game-agnostic Convergence

- One desired property of independent learning dynamics: It is game type-agnostic

Fictitious-play Property and Game-agnostic Convergence

- One desired property of independent learning dynamics: It is game type-agnostic
- Recall fictitious-play [Brown, '51]

Belief Update: For agent i maintains belief $\hat{\pi}_k^{-i}$ at time k,

$$\hat{\pi}_{k+1}^{-i} = \hat{\pi}_k^{-i} + \frac{1}{k} \cdot (a_k^{-i} - \hat{\pi}_k^{-i}),$$

Action Selection: The action a_k^i is taken from best-response

$$a_k^i \in \arg\max_{a^i} \left\{ (a^i)^T Q^i \hat{\pi}_k^{-i} \right\}.$$

Fictitious-play Property and Game-agnostic Convergence

- One desired property of independent learning dynamics: It is game type-agnostic
- Recall fictitious-play [Brown, '51]

Belief Update: For agent i maintains belief $\hat{\pi}_k^{-i}$ at time k,

$$\hat{\pi}_{k+1}^{-i} = \hat{\pi}_k^{-i} + \frac{1}{k} \cdot (a_k^{-i} - \hat{\pi}_k^{-i}),$$

Action Selection: The action a_k^i is taken from best-response

$$a_k^i \in \arg\max_{a^i} \left\{ (a^i)^T Q^i \hat{\pi}_k^{-i} \right\}.$$

  - The same update rule from each agent's perspective, converges to NE in zero-sum, identical-interest, 2×2 non-zero-sum games, etc.
  - Used to be one way to justify the universality of Nash equilibrium

Fictitious-play Property and Game-agnostic Convergence

- One desired property of independent learning dynamics: It is game type-agnostic
- Recall fictitious-play [Brown, '51]

Belief Update: For agent i maintains belief $\hat{\pi}_k^{-i}$ at time k,

$$\hat{\pi}_{k+1}^{-i} = \hat{\pi}_k^{-i} + \frac{1}{k} \cdot (a_k^{-i} - \hat{\pi}_k^{-i}),$$

Action Selection: The action a_k^i is taken from best-response

$$a_k^i \in \arg\max_{a^i} \left\{ (a^i)^T Q^i \hat{\pi}_k^{-i} \right\}.$$

  - The same update rule from each agent's perspective, converges to NE in zero-sum, identical-interest, 2×2 non-zero-sum games, etc.
  - Used to be one way to justify the universality of Nash equilibrium
  - "A game has the fictitious play property (FPP) if fictitious play process converges to its equilibrium" [Monderer and Shapley, '96]

Fictitious-play Property and Game-agnostic Convergence

- One desired property of independent learning dynamics: It is game type-agnostic
- Recall fictitious-play [Brown, '51]

Belief Update: For agent i maintains belief $\hat{\pi}_k^{-i}$ at time k,

$$\hat{\pi}_{k+1}^{-i} = \hat{\pi}_k^{-i} + \frac{1}{k} \cdot (a_k^{-i} - \hat{\pi}_k^{-i}),$$

Action Selection: The action a_k^i is taken from best-response

$$a_k^i \in \arg\max_{a^i} \left\{ (a^i)^T Q^i \hat{\pi}_k^{-i} \right\}.$$

  - The same update rule from each agent's perspective, converges to NE in zero-sum, identical-interest, 2×2 non-zero-sum games, etc.
  - Used to be one way to justify the universality of Nash equilibrium
  - "A game has the fictitious play property (FPP) if fictitious play process converges to its equilibrium" [Monderer and Shapley, '96]

What about stochastic/dynamic games?

Fictitious-play Property and Game-agnostic Convergence

- FPP can appear in SGs (with two-timescale stepsizes (as our decentralized Q-learning))

Belief Update:

$$\hat{\pi}^{-i}_{s_k,k+1} = \hat{\pi}^{-i}_{s_k,k} + \alpha_{\sharp s_k}(a^{-i}_k - \hat{\pi}^{-i}_{s_k,k})$$

Q-Value Update:

$$\hat{Q}^i_{s_k,k+1}(a) = \hat{Q}^i_{s_k,k}(a) + \beta_{\sharp s_k}\left(r^i_{s_k}(a) + \gamma \sum_{s' \in S} p(s'|s_k, a)\hat{v}^i_{s',k} - \hat{Q}^i_{s_k,k}(a)\right),$$

Fictitious-play Property and Game-agnostic Convergence

- FPP can appear in SGs (with two-timescale stepsizes (as our decentralized Q-learning))

Belief Update:

$$\hat{\pi}^{-i}_{s_k,k+1} = \hat{\pi}^{-i}_{s_k,k} + \alpha_{\sharp s_k}(a^{-i}_k - \hat{\pi}^{-i}_{s_k,k})$$

Q-Value Update:

$$\hat{Q}^i_{s_k,k+1}(a) = \hat{Q}^i_{s_k,k}(a) + \beta_{\sharp s_k}\left(r^i_{s_k}(a) + \gamma \sum_{s' \in S} p(s'|s_k, a)\hat{v}^i_{s',k} - \hat{Q}^i_{s_k,k}(a)\right),$$

where the action a^i_k is taken from best-response

$$a^i_k \in \arg\max_{a^i} \left\{(a^i)^T \hat{Q}^i_{s_k,k}\hat{\pi}^{-i}_{s_k,k}\right\}.$$

Fictitious-play Property and Game-agnostic Convergence

- FPP can appear in SGs (with two-timescale stepsizes (as our decentralized Q-learning))

  Belief Update:

$$\hat{\pi}^{-i}_{s_k,k+1} = \hat{\pi}^{-i}_{s_k,k} + \alpha_{\sharp s_k}(a^{-i}_k - \hat{\pi}^{-i}_{s_k,k})$$

  Q-Value Update:

$$\hat{Q}^i_{s_k,k+1}(a) = \hat{Q}^i_{s_k,k}(a) + \beta_{\sharp s_k}\left(r^i_{s_k}(a) + \gamma\sum_{s'\in S} p(s'|s_k,a)\hat{v}^i_{s',k} - \hat{Q}^i_{s_k,k}(a)\right),$$

  where the action a^i_k is taken from best-response

$$a^i_k \in \arg\max_{a^i}\left\{(a^i)^T\hat{Q}^i_{s_k,k}\hat{\pi}^{-i}_{s_k,k}\right\}.$$

- This very same (smoothed) fictitious-play dynamics converge to Nash equilibrium for zero-sum (competitive) and n-player identical-interest (cooperative) [SZO, '22][ZSO, '23], and multi-player zero-sum stochastic games [P*Z*O, '22], i.e., they have FPP

General-sum Cases?

- Recall: for finite-horizon case, there exists an independent algorithm as V-learning that can address general-sum SGs (CE,CCE)

General-sum Cases?

- Recall: for finite-horizon case, there exists an independent algorithm as V-learning that can address general-sum SGs (CE,CCE)
- Recall: for infinite-horizon case, value(-iteration) based approaches cannot find stationary equilibrium in general – the "NoSDE" games [Zinkevich, Greenwald, Littman, '05]

General-sum Cases?

- Recall: for finite-horizon case, there exists an independent algorithm as V-learning that can address general-sum SGs (CE,CCE)
- Recall: for infinite-horizon case, value(-iteration) based approaches cannot find stationary equilibrium in general – the "NoSDE" games [Zinkevich, Greenwald, Littman, '05]
- Our decentralized-Q learning cannot, either, as its convergence relies on the minimax value iteration (γ-contracting property of the operator) (breaks in the general-sum case)
  - It is unclear how to construct stationary equilibrium from non-stationary ones (cannot simply truncate and pick the strategy at $h = 1$, as in single-agent RL)

General-sum Cases?

- Recall: for finite-horizon case, there exists an independent algorithm as V-learning that can address general-sum SGs (CE,CCE)
- Recall: for infinite-horizon case, value(-iteration) based approaches cannot find stationary equilibrium in general – the "NoSDE" games [Zinkevich, Greenwald, Littman, '05]
- Our decentralized-Q learning cannot, either, as its convergence relies on the minimax value iteration (γ-contracting property of the operator) (breaks in the general-sum case)
  - It is unclear how to construct stationary equilibrium from non-stationary ones (cannot simply truncate and pick the strategy at $h = 1$, as in single-agent RL)

Is there a fundamental reason why infinite-horizon general-sum SGs are challenging? Is there a unique challenge compared to the finite-horizon case?

General-sum Cases?

- There might be one

Theorem (DGZ, '23)

For some constant $\epsilon > 0$, computing ϵ-(perfect) stationary CCE in 2-player stochastic games with discount factor $\gamma = 1/2$ is PPAD-hard.

- Believed to be intractable computationally [Papadimitriou, '94]
- Even 2-player and $\gamma = 1/2$, i.e., two stages in expectation

General-sum Cases?

- There might be one

Theorem (DGZ, '23)

For some constant $\epsilon > 0$, computing ϵ-(perfect) stationary CCE in 2-player stochastic games with discount factor $\gamma = 1/2$ is PPAD-hard.

- Believed to be intractable computationally [Papadimitriou, '94]
- Even 2-player and $\gamma = 1/2$, i.e., two stages in expectation
- This is in stark contrast to normal-form/static games, where CCE is tractable – showing the unique challenge in sequential and strategic decision-making
- Concurrent work (Jin et al., 2023b): similar hardness with $|\mathcal{S}|$-agents

General-sum Cases?

- There might be one

Theorem (DGZ, '23)

For some constant $\epsilon > 0$, computing ϵ-(perfect) stationary CCE in 2-player stochastic games with discount factor $\gamma = 1/2$ is PPAD-hard.

- Believed to be intractable computationally [Papadimitriou, '94]
- Even 2-player and $\gamma = 1/2$, i.e., two stages in expectation
- This is in stark contrast to normal-form/static games, where CCE is tractable – showing the unique challenge in sequential and strategic decision-making
- Concurrent work (Jin et al., 2023b): similar hardness with $|\mathcal{S}|$-agents
- Relaxing the stationary requirement enables a decentralized learning algorithm SPoCMAR with polynomial sample complexity (including the number of agents) to output a Markov equilibrium [DGZ, '23]
  - "Break the curse of multi-agents" with Markov equilibrium output

Other Independent Learning Dynamics/Algorithms?

Other Independent Learning Dynamics/Algorithms?

- All independent policy gradient methods!
  - Also referred to as "gradient play" (Shamma and Arslan, 2005), a kind of better response (as opposed to best-response)
  - Especially for Markov potential games as vanilla independent and symmetric PG simply works (Leonardos et al., 2022; Zhang et al., 2024a; Fox et al., 2022) [DWZJ, '22]

Other Independent Learning Dynamics/Algorithms?

- All independent policy gradient methods!
  - Also referred to as "gradient play" (Shamma and Arslan, 2005), a kind of better response (as opposed to best-response)
  - Especially for Markov potential games as vanilla independent and symmetric PG simply works (Leonardos et al., 2022; Zhang et al., 2024a; Fox et al., 2022) [DWZJ, '22]
  - The smooth critic variant has game-agnostic convergence to NE (zero-sum and identical-interest) (Wei et al., 2021), [DWZJ, '22]

Other Independent Learning Dynamics/Algorithms?

- All independent policy gradient methods!
  - Also referred to as "gradient play" (Shamma and Arslan, 2005), a kind of better response (as opposed to best-response)
  - Especially for Markov potential games as vanilla independent and symmetric PG simply works (Leonardos et al., 2022; Zhang et al., 2024a; Fox et al., 2022) [DWZJ, '22]
  - The smooth critic variant has game-agnostic convergence to NE (zero-sum and identical-interest) (Wei et al., 2021), [DWZJ, '22]
  - (Giannou et al., 2022): the long-run (local) behaviors of (symmetric) independent policy gradient for general-sum stochastic games

Other Independent Learning Dynamics/Algorithms?

- All independent policy gradient methods!
  - Also referred to as "gradient play" (Shamma and Arslan, 2005), a kind of better response (as opposed to best-response)
  - Especially for Markov potential games as vanilla independent and symmetric PG simply works (Leonardos et al., 2022; Zhang et al., 2024a; Fox et al., 2022) [DWZJ, '22]
  - The smooth critic variant has game-agnostic convergence to NE (zero-sum and identical-interest) (Wei et al., 2021), [DWZJ, '22]
  - (Giannou et al., 2022): the long-run (local) behaviors of (symmetric) independent policy gradient for general-sum stochastic games
- For finite-horizon setting: V-learning (Jin et al., 2023a; Song et al., 2022; Mao and Başar, 2022)

Other Independent Learning Dynamics/Algorithms?

- Another large class of independent learning algorithms: no-regret learning (in the adversarial sense)

Other Independent Learning Dynamics/Algorithms?

- Another large class of independent learning algorithms: no-regret learning (in the adversarial sense)

$$\texttt{Regret}(\pi_1, \cdots, \pi_K; K) := \max_{\pi^i \in \Pi^i} \sum_{k=1}^{K} \left(V^{i,1}_{\pi^i \times \pi_k^{-i}} - V^{i,1}_{\pi_k} \right)$$

for any sequence of product policies $\{\pi_1, \cdots, \pi_K\}$

Other Independent Learning Dynamics/Algorithms?

- Another large class of independent learning algorithms: no-regret learning (in the adversarial sense)

$$\mathtt{Regret}(\pi_1, \cdots, \pi_K; K) := \max_{\pi^i \in \Pi^i} \sum_{k=1}^{K} \left(V^{i,1}_{\pi^i \times \pi_k^{-i}} - V^{i,1}_{\pi_k} \right)$$

  for any sequence of product policies $\{\pi_1, \cdots, \pi_K\}$

  - If it can be made small (i.e., $\epsilon \cdot K$), then a uniform average of $\bar{\pi} := \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}_{\pi_k}$ is an ϵ-CCE

Other Independent Learning Dynamics/Algorithms?

- Another large class of independent learning algorithms: no-regret learning (in the adversarial sense)

$$\texttt{Regret}(\pi_1, \cdots, \pi_K; K) := \max_{\pi^i \in \Pi^i} \sum_{k=1}^{K} \left(V^{i,1}_{\pi^i \times \pi_k^{-i}} - V^{i,1}_{\pi_k} \right)$$

  for any sequence of product policies $\{\pi_1, \cdots, \pi_K\}$

  - If it can be made small (i.e., $\epsilon \cdot K$), then a uniform average of $\bar{\pi} := \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}_{\pi_k}$ is an ϵ-CCE
  - However, it is both statistically (Kwon et al., 2021; Liu et al., 2022b) and computationally (Abbasi Yadkori et al., 2013; Radanovic et al., 2019; Bai et al., 2020) intractable in general to achieve no-regret
  - Intractable even when all agents independently run an algorithm (not arbitrarily adversarial) (Foster et al., 2023b)

Other Independent Learning Dynamics/Algorithms?

- Another large class of independent learning algorithms: no-regret learning (in the adversarial sense)

$$\texttt{Regret}(\pi_1, \cdots, \pi_K; K) := \max_{\pi^i \in \Pi^i} \sum_{k=1}^{K} \left(V^{i,1}_{\pi^i \times \pi_k^{-i}} - V^{i,1}_{\pi_k} \right)$$

  for any sequence of product policies $\{\pi_1, \cdots, \pi_K\}$
  - If it can be made small (i.e., $\epsilon \cdot K$), then a uniform average of $\bar{\pi} := \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}_{\pi_k}$ is an ϵ-CCE
  - However, it is both statistically (Kwon et al., 2021; Liu et al., 2022b) and computationally (Abbasi Yadkori et al., 2013; Radanovic et al., 2019; Bai et al., 2020) intractable in general to achieve no-regret
  - Intractable even when all agents independently run an algorithm (not arbitrarily adversarial) (Foster et al., 2023b)
  - Possible if revealing the opponents' policy in the end of each episode (Liu et al., 2022b; Zhan et al., 2023) or Π^i restricted to a Markov policy class (Erez et al., 2022)

Concluding Remarks

Summary

- Multi-agent RL (theory) has expanded significantly in recent years (though we haven't really fully understood the success of AlphaGo)
- Mostly regarding (efficient) learning of stochastic games (Shapley, 1953; Fink et al., 1964; Takahashi, 1964)

Summary

- Multi-agent RL (theory) has expanded significantly in recent years (though we haven't really fully understood the success of AlphaGo)
- Mostly regarding (efficient) learning of stochastic games (Shapley, 1953; Fink et al., 1964; Takahashi, 1964)
- Classical algorithms with asymptotic convergence guarantees
- Modern algorithms with new (and mostly) non-asymptotic guarantees
  - Simulator setting
  - Online (exploration) setting
  - Offline setting
  - Computational complexities
  - Policy gradient/optimization methods

Summary

- Multi-agent RL (theory) has expanded significantly in recent years (though we haven't really fully understood the success of AlphaGo)
- Mostly regarding (efficient) learning of stochastic games (Shapley, 1953; Fink et al., 1964; Takahashi, 1964)
- Classical algorithms with asymptotic convergence guarantees
- Modern algorithms with new (and mostly) non-asymptotic guarantees
  - Simulator setting
  - Online (exploration) setting
  - Offline setting
  - Computational complexities
  - Policy gradient/optimization methods
- New models beyond (canonical) stochastic games

Summary

- (Infinite-horizon) stochastic games could have fundamental differences from matrix/static and finite-horizon games, and (infinite-horizon) MDPs/single-agent RL

Summary

- (Infinite-horizon) stochastic games could have fundamental differences from matrix/static and finite-horizon games, and (infinite-horizon) MDPs/single-agent RL
- Unique challenges compared to single-agent RL
  - Non-stationarity due to other agents' adaptation
  - "Curse of multi-agents"
  - Data coverage requirement in offline setting
  - Finite- v.s. inf-horizon & non-stationary v.s. stationary solutions
  - Computational barriers (e.g., stationary CCE, mean-field equilibrium, multi-sided nonconvexity in policy optimization)

Summary

- (Infinite-horizon) stochastic games could have fundamental differences from matrix/static and finite-horizon games, and (infinite-horizon) MDPs/single-agent RL
- Unique challenges compared to single-agent RL
  - Non-stationarity due to other agents' adaptation
  - "Curse of multi-agents"
  - Data coverage requirement in offline setting
  - Finite- v.s. inf-horizon & non-stationary v.s. stationary solutions
  - Computational barriers (e.g., stationary CCE, mean-field equilibrium, multi-sided nonconvexity in policy optimization)
- Unique challenges compared to learning in matrix/static games
  - New scenarios: simulator, online exploration, offline learning
  - Markovian sampling
  - Computational barriers (e.g., stationary CCE, nonconvexity in policy optimization)
  - Hardness of no-regret learning

Summary

- (Infinite-horizon) stochastic games could have fundamental differences from matrix/static and finite-horizon games, and (infinite-horizon) MDPs/single-agent RL
- Unique challenges compared to single-agent RL
  - Non-stationarity due to other agents' adaptation
  - "Curse of multi-agents"
  - Data coverage requirement in offline setting
  - Finite- v.s. inf-horizon & non-stationary v.s. stationary solutions
  - Computational barriers (e.g., stationary CCE, mean-field equilibrium, multi-sided nonconvexity in policy optimization)
- Unique challenges compared to learning in matrix/static games
  - New scenarios: simulator, online exploration, offline learning
  - Markovian sampling
  - Computational barriers (e.g., stationary CCE, nonconvexity in policy optimization)
  - Hardness of no-regret learning
  - Function approximation (did not cover much here)
  - Partial observations (did not cover much here)

Additional Thoughts

- If multi-agent RL is the answer, justifying equilibrium as the naturally emerging behavior of independent and natural adaptation/learning dynamics, and studying their long-run behaviors, might be some questions (among many other significant ones, e.g., sample and computational complexities, regret, convergence rates, etc.)

Additional Thoughts

- If multi-agent RL is the answer, justifying equilibrium as the naturally emerging behavior of independent and natural adaptation/learning dynamics, and studying their long-run behaviors, might be some questions (among many other significant ones, e.g., sample and computational complexities, regret, convergence rates, etc.)
  - Independent learning dynamics can be made simple for SGs
  - Fictitious-play property and game-agnostic convergence can exist

Additional Thoughts

- If multi-agent RL is the answer, justifying equilibrium as the naturally emerging behavior of independent and natural adaptation/learning dynamics, and studying their long-run behaviors, might be some questions (among many other significant ones, e.g., sample and computational complexities, regret, convergence rates, etc.)
  - Independent learning dynamics can be made simple for SGs
  - Fictitious-play property and game-agnostic convergence can exist

Thank You!

References

Abbasi Yadkori, Y., Bartlett, P. L., Kanade, V., Seldin, Y. and Szepesvári, C. (2013). Online learning in Markov decision processes with adversarially chosen transition probability distributions. *Advances in Neural Information Processing Systems*, **26**.

Agarwal, A., Kakade, S. M., Lee, J. D. and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, **22** 1–76.

Aggarwal, S., Bastopcu, M., Başar, T. et al. (2024). Policy optimization finds nash equilibrium in regularized general-sum LQ games. *arXiv preprint arXiv:2404.00045*.

Alacaoglu, A., Viano, L., He, N. and Cevher, V. (2022). A natural actor-critic framework for zero-sum Markov games. In *International Conference on Machine Learning*. PMLR.

Amazon (2023). Warehouse mobile robots.

Anagnostides, I., Panageas, I., Farina, G. and Sandholm, T. (2024). Optimistic policy gradient in multi-player Markov games with a single controller: Convergence beyond the minty property. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38.

Anahtarci, B., Kariksiz, C. D. and Saldi, N. (2023). Q-learning in regularized mean-field games. *Dynamic Games and Applications*, **13** 89–117.

Arslan, G. and Yüksel, S. (2017). Decentralized Q-learning for stochastic teams and games. *IEEE Transactions on Automatic Control*, **62** 1545–1558.

Aydin, S. and Eksin, C. (2023). Networked policy gradient play in Markov potential games. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Bai, Y. and Jin, C. (2020). Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*.

Bai, Y., Jin, C. and Yu, T. (2020). Near-optimal reinforcement learning with self-play. *Advances in Neural Information Processing Systems*, **33**.

Başar, T. and Bernhard, P. (1995). *H_∞ Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Birkhäuser, Boston.

Başar, T. and Olsder, G. J. (1998). *Dynamic Noncooperative Game Theory*. SIAM.

Beck, C. L. and Srikant, R. (2012). Error bounds for constant step-size Q-learning. *Systems & Control Letters*, **61** 1203–1208.

Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C. et al. (2019). Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*.

Bertsekas, D. (2021). Distributed asynchronous policy iteration for sequential zero-sum games and minimax control. *arXiv preprint arXiv:2107.10406*.

Bhandari, J. and Russo, D. (2024). Global optimality guarantees for policy gradient methods. *Operations Research*.

Bowling, M. (2000). Convergence problems of general-sum multiagent reinforcement learning. In *ICML*.

Bowling, M. and Veloso, M. (2001). Rational and convergent learning in stochastic games. In *International Joint Conference on Artificial Intelligence*, vol. 17.

Boyd, S., El Ghaoui, L., Feron, E. and Balakrishnan, V. (1994). *Linear matrix inequalities in system and control theory*. SIAM.

Brafman, R. I. and Tennenholtz, M. (2000). A near-optimal polynomial time algorithm for learning in certain classes of stochastic games. *Artificial Intelligence*, **121** 31–47.

Brafman, R. I. and Tennenholtz, M. (2002). R-MAX-A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, **3** 213–231.

Brahma, S., Bai, Y., Do, D. A. and Doan, T. T. (2022). Convergence rates of asynchronous policy iteration for zero-sum Markov games under stochastic and optimistic settings. In *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE.

Bu, J., Ratliff, L. J. and Mesbahi, M. (2019). Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv preprint arXiv:1911.04672*.

Bubeck, S., Li, Y., Peres, Y. and Sellke, M. (2020). Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without. In *Conference on Learning Theory*. PMLR.

Cai, Q., Yang, Z., Jin, C. and Wang, Z. (2020). Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*. PMLR.

Cai, Y., Candogan, O., Daskalakis, C. and Papadimitriou, C. (2016). Zero-sum polymatrix games: A generalization of minmax. *Mathematics of Operations Research*, **41** 648–655.

Cai, Y., Luo, H., Wei, C.-Y. and Zheng, W. (2024a). Near-optimal policy optimization for correlated equilibrium in general-sum Markov games. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Cai, Y., Luo, H., Wei, C.-Y. and Zheng, W. (2024b). Uncoupled and convergent learning in two-player zero-sum Markov games with bandit feedback. *Advances in Neural Information Processing Systems*, **36**.

Cen, S., Cheng, C., Chen, Y., Wei, Y. and Chi, Y. (2022). Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, **70** 2563–2578.

Cen, S., Chi, Y., Du, S. and Xiao, L. (2023). Faster last-iterate convergence of policy optimization in zero-sum Markov games. In *International Conference on Learning Representations (ICLR)*.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*. PMLR.

Chen, X., Deng, X. and Teng, S.-H. (2009). Settling the complexity of computing two-player Nash equilibria. *Journal of the ACM*, **56** 14.

Chen, X., Qu, G., Tang, Y., Low, S. and Li, N. (2022a). Reinforcement learning for selective key applications in power systems: Recent advances and future challenges. *IEEE Transactions on Smart Grid*, **13** 2935–2958.

Chen, Z., Ma, S. and Zhou, Y. (2022b). Sample efficient stochastic policy extragradient algorithm for zero-sum Markov game. In *International Conference on Learning Representation*.

Chen, Z., Zhou, D. and Gu, Q. (2022c). Almost optimal algorithms for two-player zero-sum linear mixture Markov games. In *International Conference on Algorithmic Learning Theory*. PMLR.

Cheng, M., Zhou, R., Kumar, P. and Tian, C. (2024). Provable policy gradient methods for average-reward Markov potential games. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Chu, T., Chinchali, S. and Katti, S. (2019). Multi-agent reinforcement learning for networked system control. In *International Conference on Learning Representations*.

Condon, A. (1990). On algorithms for simple stochastic games. *Advances in Computational Complexity Theory*, **13** 51–72.

Cui, K. and Koeppl, H. (2021). Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Cui, Q. and Du, S. S. (2022a). Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. *Advances in Neural Information Processing Systems*, **35** 11739–11751.

Cui, Q. and Du, S. S. (2022b). When are offline two-player zero-sum Markov games solvable? *Advances in Neural Information Processing Systems*, **35** 25779–25791.

Daskalakis, C., Foster, D. J. and Golowich, N. (2020). Independent policy gradient methods for competitive reinforcement learning. In *Advances in Neural Information Processing Systems*.

Daskalakis, C., Goldberg, P. W. and Papadimitriou, C. H. (2009). The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, **39** 195–259.

Doan, T., Maguluri, S. and Romberg, J. (2019). Finite-time analysis of distributed TD (0) with linear function approximation on multi-agent reinforcement learning. In *International Conference on Machine Learning*.

Doan, T. T., Maguluri, S. T. and Romberg, J. (2021). Finite-time performance of distributed temporal-difference learning with linear function approximation. *SIAM Journal on Mathematics of Data Science*, **3** 298–320.

Duchi, J. C., Agarwal, A. and Wainwright, M. J. (2011). Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, **57** 592–606.

Erez, L., Lancewicki, T., Sherman, U., Koren, T. and Mansour, Y. (2022). Regret minimization and convergence to equilibria in general-sum Markov games. *arXiv preprint arXiv:2207.14211*.

Erez, L., Lancewicki, T., Sherman, U., Koren, T. and Mansour, Y. (2023). Regret minimization and convergence to equilibria in general-sum Markov games. In *International Conference on Machine Learning*. PMLR.

Even-Dar, E., Mansour, Y. and Bartlett, P. (2003). Learning rates for Q-learning. *Journal of Machine Learning Research*, **5**.

FAIR (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, **378** 1067–1074.

Fatkhullin, I., Barakat, A., Kireeva, A. and He, N. (2023). Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. In *International Conference on Machine Learning*. PMLR.

Figura, M., Kosaraju, K. C. and Gupta, V. (2021). Adversarial attacks in consensus-based multi-agent reinforcement learning. In *2021 American control conference (ACC)*. IEEE.

Filar, J. and Vrieze, K. (2012). *Competitive Markov Decision Processes*. Springer Science & Business Media.

Filar, J. A. and Tolwinski, B. (1991). On the algorithm of Pollatschek and Avi-Itzhak. In *Stochastic Games And Related Topics: In Honor of Professor LS Shapley*. Springer, 59–70.

Fink, A. M. et al. (1964). Equilibrium in a stochastic n-person game. *Journal of Science of the Hiroshima University, Series A-I (Mathematics)*, **28** 89–93.

Flesch, J., Thuijsman, F. and Vrieze, O. J. (2007). Stochastic games with additive transitions. *European Journal of Operational Research*, **179** 483–497.

Foster, D., Foster, D. J., Golowich, N. and Rakhlin, A. (2023a). On the complexity of multi-agent decision making: From learning in games to partial monitoring. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR.

Foster, D. J., Golowich, N. and Kakade, S. M. (2023b). Hardness of independent learning and sparse equilibrium computation in Markov games. In *International Conference on Machine Learning*. PMLR.

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*. PMLR.

Gao, Z., Ma, Q., Başar, T. and Birge, J. R. (2021). Finite-sample analysis of decentralized Q-learning for stochastic games. *arXiv preprint arXiv:2112.07859*.

Geist, M., Pérolat, J., Laurière, M., Elie, R., Perrin, S., Bachem, O., Munos, R. and Pietquin, O. (2022). Concave utility reinforcement learning: The mean-field game viewpoint. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*.

Giannou, A., Lotidis, K., Mertikopoulos, P. and Vlatakis-Gkaragkounis, E.-V. (2022). On the convergence of policy gradient methods to Nash equilibria in general stochastic games. *Advances in Neural Information Processing Systems*, **35** 7128–7141.

Greenwald, A., Hall, K. and Serrano, R. (2003). Correlated Q-learning. In *International Conference on Machine Learning*, vol. 20.

Guo, H., Fu, Z., Yang, Z. and Wang, Z. (2021). Decentralized single-timescale actor-critic on zero-sum two-player stochastic games. In *International Conference on Machine Learning*. PMLR.

Guo, X., Hu, A., Xu, R. and Zhang, J. (2019). Learning mean-field games. *Advances in Neural Information Processing Systems*, **32**.

Guo, X., Hu, A., Xu, R. and Zhang, J. (2023a). A general framework for learning mean-field games. *Mathematics of Operations Research*, **48** 656–686.

Guo, X., Li, X., Maheshwari, C., Sastry, S. and Wu, M. (2023b). Markov α-potential games. *arXiv preprint arXiv:2305.12553*.

Hambly, B., Xu, R. and Yang, H. (2023). Policy gradient methods find the Nash equilibrium in n-player general-sum linear-quadratic games. *Journal of Machine Learning Research*, **24** 1–56.

Hansen, T. D., Miltersen, P. B. and Zwick, U. (2013). Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM*, **60** 1.

Heinrich, J. and Silver, D. (2016). Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*.

Hoffman, A. J. and Karp, R. M. (1966). On nonterminating stochastic games. *Management Science*, **12** 359–370.

Hosseinirad, S., Salizzoni, G., Porzani, A. A. and Kamgarpour, M. (2024). On linear quadratic potential games. *arXiv:2305.13476*.

Hu, J. and Wellman, M. P. (2003). Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, **4** 1039–1069.

Huang, B., Lee, J. D., Wang, Z. and Yang, Z. (2022). Towards general function approximation in zero-sum Markov games. In *International Conference on Learning Representations*.

Huang, J., He, N. and Krause, A. (2024a). Model-based RL for mean-field games is not statistically harder than single-agent RL. In *Forty-first International Conference on Machine Learning*.

Huang, J., Yardim, B. and He, N. (2024b). On the statistical efficiency of mean-field reinforcement learning with general function approximation. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

IBM (1997). Deepblue.

Jaakkola, T., Jordan, M. and Singh, S. (1993). Convergence of stochastic iterative dynamic programming algorithms. *Advances in Neural Information Processing Systems*, **6**.

Jaksch, T., Ortner, R. and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, **11** 1563–1600.

JIN, C., LIU, Q., WANG, Y. and YU, T. (2023a). V-learning – a simple, efficient, decentralized algorithm for multiagent reinforcement learning. *Mathematics of Operations Research*.

JIN, C., LIU, Q. and YU, T. (2022). The power of exploiter: Provable multi-agent RL in large state spaces. In *International Conference on Machine Learning*. PMLR.

JIN, Y., MUTHUKUMAR, V. and SIDFORD, A. (2023b). The complexity of infinite-horizon general-sum stochastic games. In *Innovations in Theoretical Computer Science Conference (ITCS 2023)*.

JIN, Y., YANG, Z. and WANG, Z. (2021). Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*. PMLR.

KAKADE, S. and LANGFORD, J. (2002). Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, vol. 2.

KAKADE, S. M. (2003). *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom).

KALOGIANNIS, F., ANAGNOSTIDES, I., PANAGEAS, I., VLATAKIS-GKARAGKOUNIS, E.-V., CHATZIAFRATIS, V. and STAVROULAKIS, S. A. (2023). Efficiently computing Nash equilibria in adversarial team Markov games. In *The Eleventh International Conference on Learning Representations*.

KALOGIANNIS, F. and PANAGEAS, I. (2023). Zero-sum polymatrix Markov games: Equilibrium collapse and efficient computation of nash equilibria. *Advances in Neural Information Processing Systems*, **36**.

KAR, S., MOURA, J. M. and POOR, H. V. (2013). QD-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations. *IEEE Transactions on Signal Processing*, **61** 1848–1862.

KEARNS, M. J. and SINGH, S. P. (1999). Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems*.

KURACH, K., RAICHUK, A., STAŃCZYK, P., ZAJAC, M., BACHEM, O., ESPEHOLT, L., RIQUELME, C., VINCENT, D., MICHALSKI, M., BOUSQUET, O. ET AL. (2020). Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34.

KWON, J., EFRONI, Y., CARAMANIS, C. and MANNOR, S. (2021). RL for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, **34** 24523–24534.

LANCTOT, M., LOCKHART, E., LESPIAU, J.-B., ZAMBALDI, V., UPADHYAY, S., PÉROLAT, J., SRINIVASAN, S., TIMBERS, F., TUYLS, K., OMIDSHAFIEI, S. ET AL. (2019). Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*.

LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit algorithms*. Cambridge University Press.

LEE, D. (2023). Finite-time analysis of minimax Q-learning for two-player zero-sum Markov games: Switching system approach. *arXiv preprint arXiv:2306.05700*.

Lee, D., Yoon, H. and Hovakimyan, N. (2018). Primal-dual algorithm for distributed reinforcement learning: distributed GTD. In *IEEE Conference on Decision and Control*.

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*.

Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J. and Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, **37** 421–436.

Li, G., Chi, Y., Wei, Y. and Chen, Y. (2022). Minimax-optimal multi-agent RL in Markov games with a generative model. *Advances in Neural Information Processing Systems*, **35** 15353–15367.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings*. Elsevier, 157–163.

Littman, M. L. (2001). Friend-or-foe Q-learning in general-sum games. In *International Conference on Machine Learning*, vol. 1.

Littman, M. L. and Szepesvári, C. (1996). A generalized reinforcement-learning model: Convergence and applications. In *ICML*, vol. 96.

Liu, K. and Zhao, Q. (2010). Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, **58** 5667–5681.

Liu, Q., Szepesvári, C. and Jin, C. (2022a). Sample-efficient reinforcement learning of partially observable Markov games. *Advances in Neural Information Processing Systems*, **35** 18296–18308.

Liu, Q., Wang, Y. and Jin, C. (2022b). Learning Markov games with adversarial opponents: Efficient algorithms and fundamental limits. In *International Conference on Machine Learning*. PMLR.

Liu, Q., Yu, T., Bai, Y. and Jin, C. (2021). A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*. PMLR.

Liu, X.-Y., Xia, Z., Rui, J., Gao, J., Yang, H., Zhu, M., Wang, C. D., Wang, Z. and Guo, J. (2022c). Finrl-Meta: Market environments and benchmarks for data-driven financial reinforcement learning. *NeurIPS*.

Liu, Z., Lu, M., Xiong, W., Zhong, H., Hu, H., Zhang, S., Zheng, S., Yang, Z. and Wang, Z. (2024). Maximize to explore: One objective function fusing estimation, planning, and exploration. *Advances in Neural Information Processing Systems*, **36**.

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O. and Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, **30**.

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

MAHAJAN, A. (2008). *Sequential Decomposition of Sequential Dynamic Teams: Applications to Real-Time Communication and Networked Control Systems*. Ph.D. thesis, University of Michigan.

MAO, W. and BAŞAR, T. (2022). Provably efficient reinforcement learning in decentralized general-sum Markov games. *Dynamic Games and Applications* 1–22.

MAO, W., QIU, H., WANG, C., FRANKE, H., KALBARCZYK, Z. and BAŞAR, T. (2024). $o(1/t)$ convergence to (coarse) correlated equilibria in full-information general-sum Markov games. *arXiv preprint arXiv:2403.07890*.

MARDEN, J. R. (2012). State based potential games. *Automatica*, **48** 3075–3088.

MAZUMDAR, E., RATLIFF, L. J., JORDAN, M. I. and SASTRY, S. S. (2019). Policy-gradient algorithms have no guarantees of convergence in continuous action and state multi-agent settings. *arXiv preprint arXiv:1907.03712*.

MAZUMDAR, E., RATLIFF, L. J., JORDAN, M. I. and SASTRY, S. S. (2020). Policy-gradient algorithms have no guarantees of convergence in linear quadratic games. In *AAMAS Conference proceedings*.

MUNOS, R. and SZEPESVÁRI, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, **9** 815–857.

NAYYAR, A., GUPTA, A., LANGBORT, C. and BAŞAR, T. (2013a). Common information based Markov perfect equilibria for stochastic games with asymmetric information: Finite games. *IEEE Transactions on Automatic Control*, **59** 555–570.

NAYYAR, A., MAHAJAN, A. and TENEKETZIS, D. (2013b). Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, **58** 1644–1658.

NEDIC, A. and OZDAGLAR, A. (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, **54** 48–61.

OLSSON, J., ZHANG, R., TEGLING, E. and LI, N. (2024). Scalable reinforcement learning for linear-quadratic control of networks. *arXiv preprint arXiv:2401.16183*.

OPENAI (2022). Chatgpt.

OUHAMMA, R. and KAMGARPOUR, M. (2023). Learning nash equilibria in zero-sum Markov games: A single time-scale algorithm under weak reachability. *arXiv preprint arXiv:2312.08008*.

PATEK, S. D. (1997). *Stochastic and Shortest Path Games: Theory and Algorithms*. Ph.D. thesis, Massachusetts Institute of Technology.

PÉROLAT, J., PERRIN, S., ELIE, R., LAURIÈRE, M., PILIOURAS, G., GEIST, M., TUYLS, K. and PIETQUIN, O. (2022). Scaling mean field games by online mirror descent. In *International Conference on Autonomous Agents and Multiagent Systems*.

PERRIN, S., PÉROLAT, J., LAURIÈRE, M., GEIST, M., ELIE, R. and PIETQUIN, O. (2020). Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems*, **33** 13199–13213.

Pollatschek, M. and Avi-Itzhak, B. (1969). Algorithms for stochastic games with geometrical interpretation. *Management Science*, **15** 399–415.

Qiu, S., Dai, Z., Zhong, H., Wang, Z., Yang, Z. and Zhang, T. (2024). Posterior sampling for competitive RL: Function approximation and partial observation. *Advances in Neural Information Processing Systems*, **36**.

Qu, G., Lin, Y., Wierman, A. and Li, N. (2020). Scalable multi-agent reinforcement learning for networked systems with average reward. *Advances in Neural Information Processing Systems*, **33** 2074–2086.

Qu, G., Wierman, A. and Li, N. (2022). Scalable reinforcement learning for multiagent networked systems. *Operations Research*, **70** 3601–3628.

Radanovic, G., Devidze, R., Parkes, D. and Singla, A. (2019). Learning to collaborate in Markov decision processes. In *International Conference on Machine Learning*. PMLR.

Ramponi, G., Kolev, P., Pietquin, O., He, N., Laurière, M. and Geist, M. (2024). On imitation in mean-field games. *Advances in Neural Information Processing Systems*, **36**.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J. and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, **34** 11702–11716.

Samvelyan, M., Rashid, T., Schroeder de Witt, C., Farquhar, G., Nardelli, N., Rudner, T. G., Hung, C.-M., Torr, P. H., Foerster, J. and Whiteson, S. (2019). The StarCraft multi-agent challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shalev-Shwartz, S., Shammah, S. and Shashua, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*.

Shamma, J. S. and Arslan, G. (2005). Dynamic fictitious play, dynamic gradient play, and distributed convergence to Nash equilibria. *IEEE Transactions on Automatic Control*, **50** 312–327.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences*, **39** 1095–1100.

Sidford, A., Wang, M., Yang, L. and Ye, Y. (2020). Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, **529** 484–489.

Song, Z., Lee, J. D. and Yang, Z. (2023). Can we find Nash equilibria at a linear rate in Markov games? In *The Eleventh International Conference on Learning Representations*.

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*.

Stanković, M. S., Beko, M. and Stanković, S. S. (2023). Distributed consensus-based multi-agent temporal-difference learning. *Automatica*, **151** 110922.

Subramanian, J., Sinha, A. and Mahajan, A. (2023). Robustness and sample complexity of model-based MARL for general-sum Markov games. *Dynamic Games and Applications*, **13** 56–88.

Sun, J., Wang, G., Giannakis, G. B., Yang, Q. and Yang, Z. (2020). Finite-time analysis of decentralized temporal-difference learning with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Sun, Y., Liu, T., Kumar, P. and Shahrampour, S. (2024). Linear convergence of independent natural policy gradient in games with entropy regularization. *IEEE Control Systems Letters*.

Sun, Y., Liu, T., Zhou, R., Kumar, P. and Shahrampour, S. (2023). Provably fast convergence of independent natural policy gradient for Markov potential games. *Advances in Neural Information Processing Systems*, **36** 43951–43971.

Sutton, R. S., McAllester, D. A., Singh, S. P. and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*.

Szepesvári, C. (2022). *Algorithms for reinforcement learning*. Springer Nature.

Szepesvári, C. and Littman, M. L. (1999). A unified analysis of value-function-based reinforcement-learning algorithms. *Neural Computation*, **11** 2017–2060.

Takahashi, M. (1964). Equilibrium points of stochastic non-cooperative n-person games. *Journal of Science of the Hiroshima University, Series AI (Mathematics)*, **28** 95–99.

Terry, J., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L. S., Dieffendahl, C., Horsch, C., Perez-Vicente, R. et al. (2021). Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, **34** 15032–15043.

Tsitsiklis, J. and Athans, M. (1985). On the complexity of decentralized decision making and detection problems. *IEEE Transactions on Automatic Control*, **30** 440–446.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning*, **16** 185–202.

Van Der Wal, J. (1978). Discounted markov games: Generalized policy iteration method. *Journal of Optimization Theory and Applications*, **25** 125–138.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P. et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, **575** 350–354.

Wai, H.-T., Yang, Z., Wang, Z. and Hong, M. (2018). Multi-agent reinforcement learning via double averaging primal-dual optimization. In *Advances in Neural Information Processing Systems*.

Wang, L., Cai, Q., Yang, Z. and Wang, Z. (2020). Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*.

Wang, Y., Liu, Q., Bai, Y. and Jin, C. (2023). Breaking the curse of multiagency: Provably efficient decentralized multi-agent RL with function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, **8** 279–292.

Wei, C.-Y., Hong, Y.-T. and Lu, C.-J. (2017). Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*.

Wei, C.-Y., Lee, C.-W., Zhang, M. and Luo, H. (2021). Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In *Conference on Learning Theory*. PMLR.

Whittle, P. (1981). Risk-sensitive linear/quadratic/Gaussian control. *Advances in Applied Probability*, **13** 764–777.

Winnicki, A. and Srikant, R. (2023). A new policy iteration algorithm for reinforcement learning in zero-sum Markov games. *arXiv preprint arXiv:2303.09716*.

Witsenhausen, H. S. (1968). A counterexample in stochastic optimum control. *SIAM Journal on Control*, **6** 131–147.

Wu, J., Barakat, A., Fatkhullin, I. and He, N. (2023). Learning zero-sum linear quadratic games with improved sample complexity. In *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE.

Xiao, L., Boyd, S. and Kim, S.-J. (2007). Distributed average consensus with least-mean-square deviation. *Journal of parallel and distributed computing*, **67** 33–46.

Xie, Q., Chen, Y., Wang, Z. and Yang, Z. (2020). Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*. PMLR.

Xie, Q., Yang, Z., Wang, Z. and Minca, A. (2021a). Learning while playing in mean-field games: Convergence and optimality. In *International Conference on Machine Learning*. PMLR.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P. and Agarwal, A. (2021b). Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, **34** 6683–6694.

Xiong, W., Zhong, H., Shi, C., Shen, C. and Zhang, T. (2022). A self-play posterior sampling algorithm for zero-sum Markov games. In *International Conference on Machine Learning*. PMLR.

Yang, Y. and Ma, C. (2023). $o(1/t)$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum Markov games. In *The Eleventh International Conference on Learning Representations*.

Yardim, B., Cayci, S., Geist, M. and He, N. (2023). Policy mirror ascent for efficient and independent learning in mean field games. In *International Conference on Machine Learning*. PMLR.

Yardim, B., Goldman, A. and He, N. (2024). When is mean-field reinforcement learning tractable and relevant? In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*.

Zeng, S., Doan, T. and Romberg, J. (2022). Regularized gradient descent ascent for two-player zero-sum Markov games. *Advances in Neural Information Processing Systems*, **35** 34546–34558.

Zhan, W., Huang, B., Huang, A., Jiang, N. and Lee, J. (2022). Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*. PMLR.

Zhan, W., Lee, J. D. and Yang, Z. (2023). Decentralized optimistic hyperpolicy mirror descent: Provably no-regret learning in Markov games. In *The Eleventh International Conference on Learning Representations*.

Zhang, R., Liu, Q., Wang, H., Xiong, C., Li, N. and Bai, Y. (2022a). Policy optimization for Markov games: Unified framework and faster convergence. *Advances in Neural Information Processing Systems*, **35** 21886–21899.

Zhang, R., Mei, J., Dai, B., Schuurmans, D. and Li, N. (2022b). 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., Ren, Z. and Li, N. (2024a). Gradient play in stochastic games: stationary points, convergence, and sample complexity. *IEEE Transactions on Automatic Control*.

Zhang, Y., Qu, G., Xu, P., Lin, Y., Chen, Z. and Wierman, A. (2023). Global convergence of localized policy iteration in networked multi-agent reinforcement learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, **7** 1–51.

Zhang, Y. and Zavlanos, M. M. (2019). Distributed off-policy actor-critic reinforcement learning with policy consensus. In *2019 IEEE 58th Conference on decision and control (CDC)*. IEEE.

Zhang, Y., Zhang, Z., Quiñones-Grueiro, M., Barbour, W., Weston, C., Biswas, G. and Work, D. (2024b). Field deployment of multi-agent reinforcement learning based variable speed limit controllers. *arXiv preprint arXiv:2407.08021*.

Zhao, Y., Tian, Y., Lee, J. and Du, S. (2022). Provably efficient policy optimization for two-player zero-sum markov games. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Zhong, H., Xiong, W., Tan, J., Wang, L., Zhang, T., Wang, Z. and Yang, Z. (2022). Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *International Conference on Machine Learning*. PMLR.

Zhou, Z., Chen, Z., Lin, Y. and Wierman, A. (2023). Convergence rates for localized actor-critic in networked Markov potential games. In *Uncertainty in Artificial Intelligence*. PMLR.

Zinkevich, M., Greenwald, A. and Littman, M. L. (2006). Cyclic equilibria in Markov games. In *Advances in Neural Information Processing Systems*.