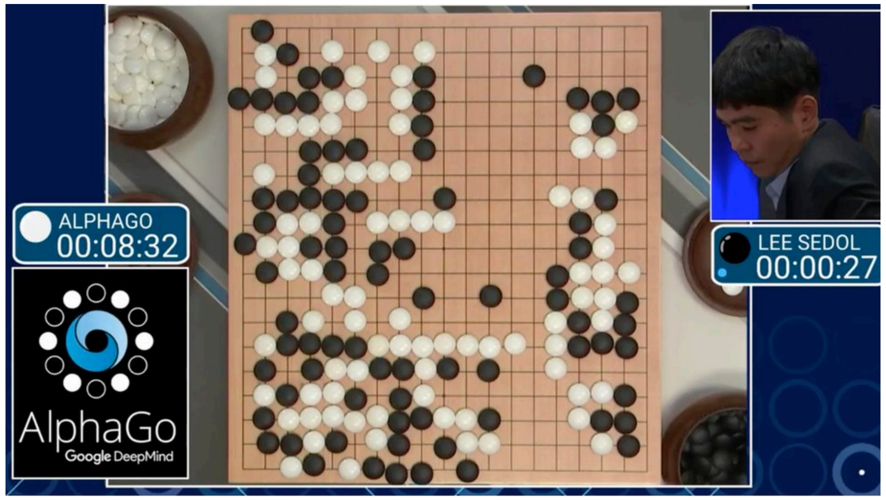
Recent Progresses on the Theory of Multi-Agent Reinforcement Learning and Games

Yu Bai
Salesforce Research

Blog post: https://yubai.org/blog/marl_theory.html

Multi-Agent Reinforcement Learning



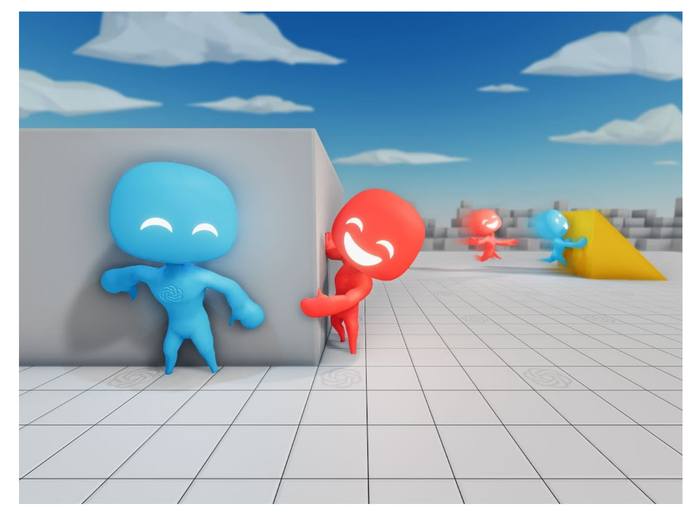
AlphaGo

Poker

AI Economist

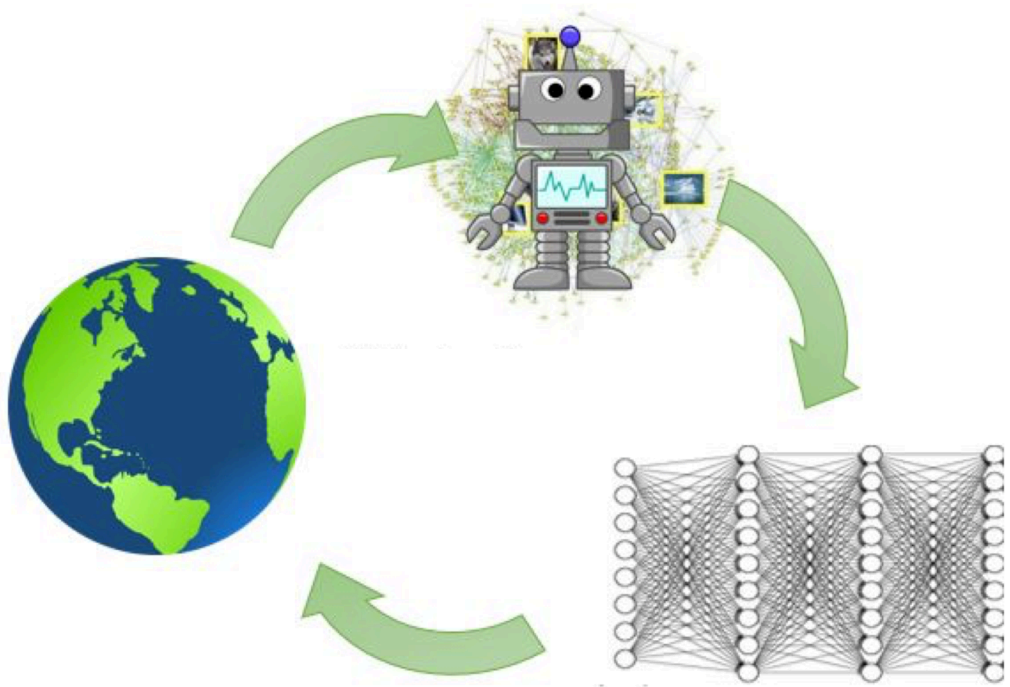
Starcraft

Diplomacy

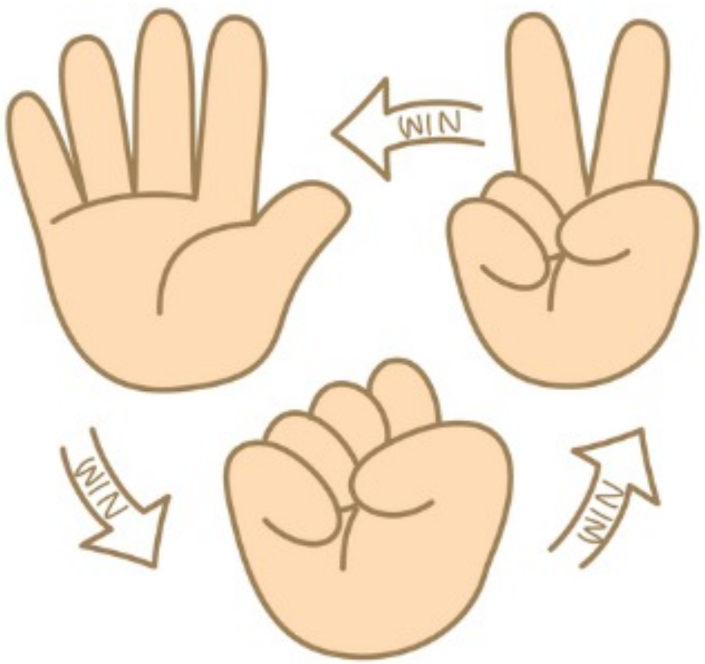


Hide and Seek

Multi-Agent Reinforcement Learning



sequential decisions



multi-agent

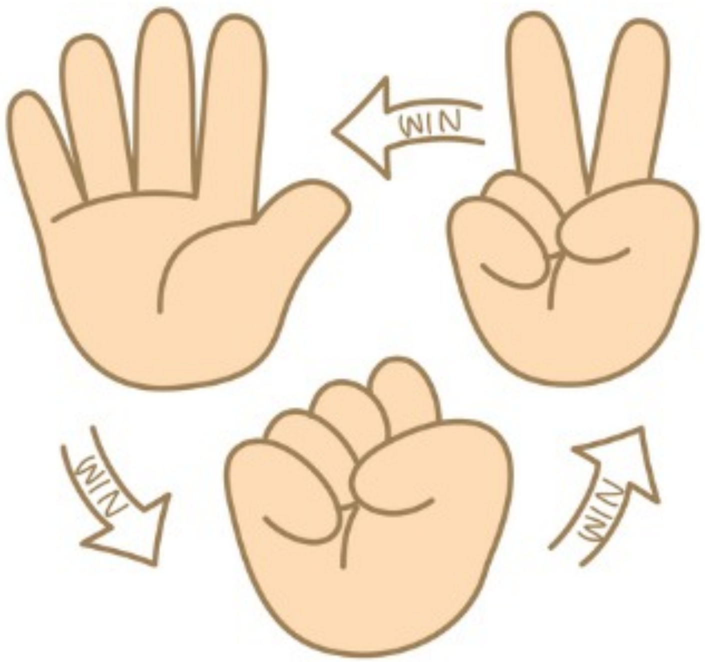
A relatively new field, with unique challenges and opportunities
for both **theory**/empirical research.

Outline

- Formulations
  - Normal-Form Games (NFGs)
  - Markov Games (MGs)
- Two-Player Zero-Sum Markov Games
- Multi-Player General-Sum Markov Games
- Faster Convergence via Optimistic Algorithms
- Advanced Topics
  - Imperfect Information
  - Rationalizability

* Sketchy -> Please refer to slides / references (in presenter notes)

Normal-Form Games (NFGs)



			
	0	-1	+1
	+1	0	-1
	-1	+1	0

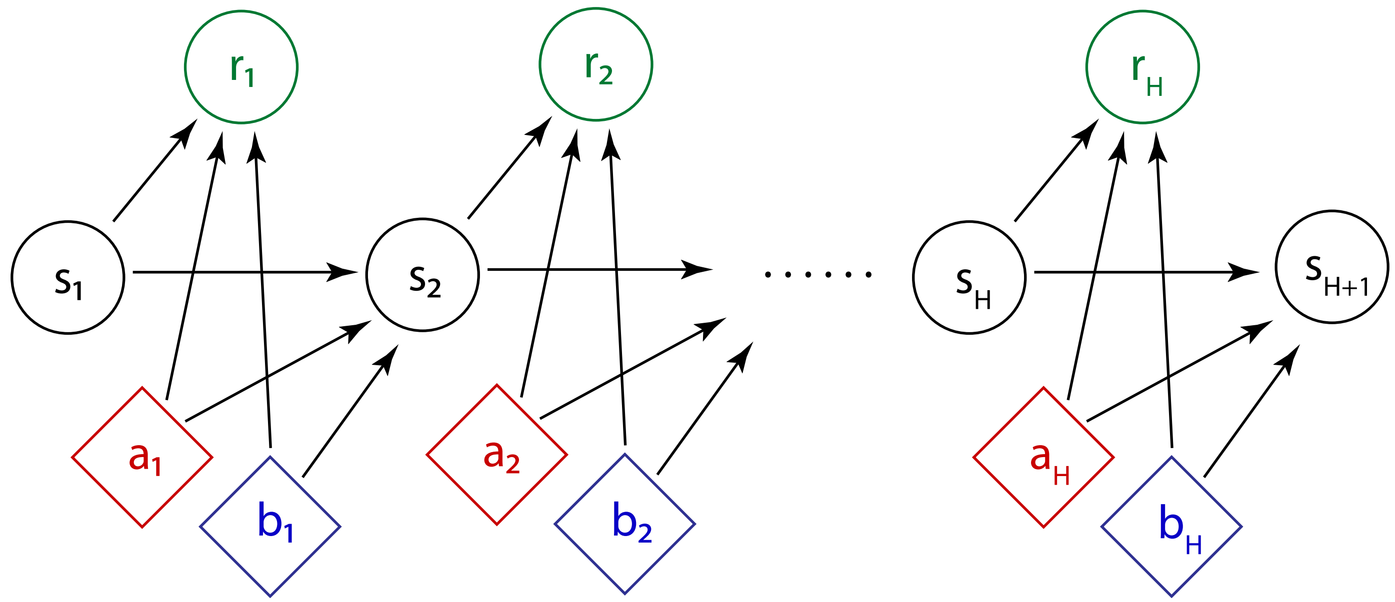
Multi-player Normal-Form Games (NFGs):

- Players $\{1,\ldots,m\}$
- Each player i chooses their action $a_i \in \mathcal{A}_i$ **simultaneously**
- Each player i receives reward $r_i(a_1,\ldots,a_m) \in [0,1]$ (**general-sum**)

Markov Games (MGs)

[Shapley 1953]

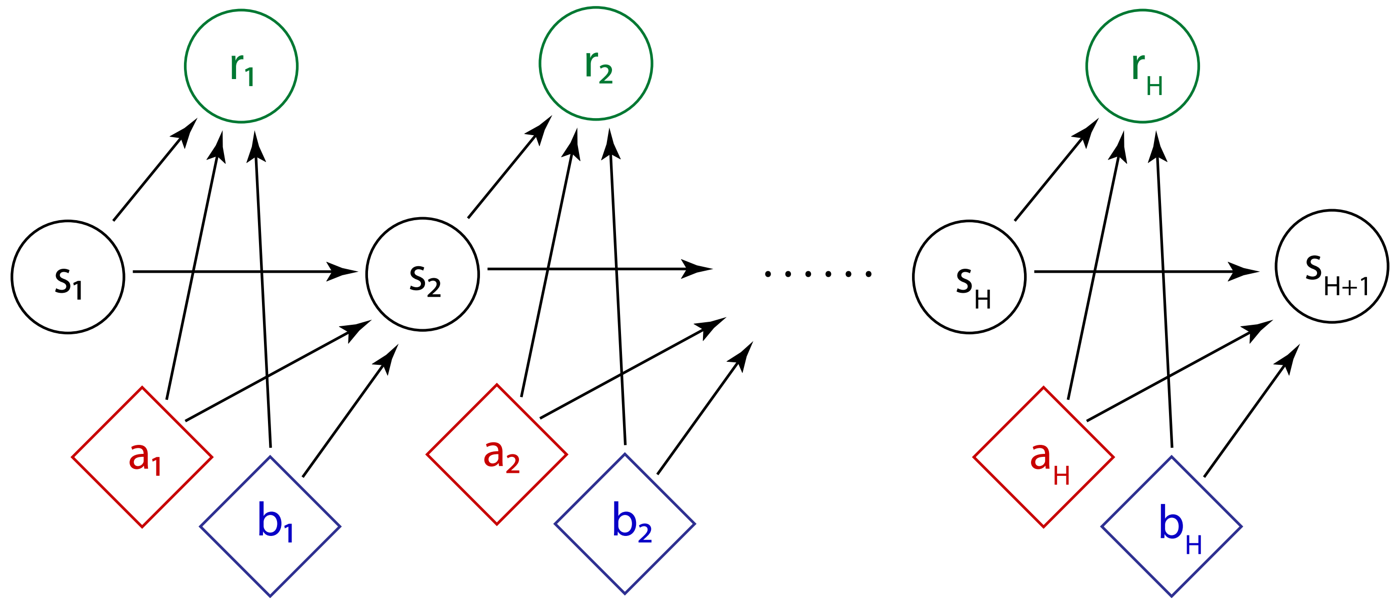
(also known as **Stochastic Games**)



Finite-horizon General-Sum Markov Games with m players:

- Horizon length H
- State space $|\mathcal{S}| = S$
- Action space $|\mathcal{A}_i| = A_i$ (for i-th player)
- Reward: $r_{i,h}(s_h, a_{1,h}, \ldots, a_{m,h})$ (for i-th player)
- Transition: $(s_h, a_{1,h}, \ldots, a_{m,h}) \rightarrow s_{h+1}$

Policies, Values, Equilibria



- (Markov product) policy: $a_{i,h} \sim \pi_{i,h}(\cdot \mid s_h)$
- Game value (for i-th player): $V_i^{\pi} = \mathbb{E}_{\pi}\left[\sum_{h=1}^{H} r_{i,h}\right]$

Nash Equilibrium (NE): A product policy $\pi = \{\pi_i\}_{i\in[m]}$ is an ε-NE if

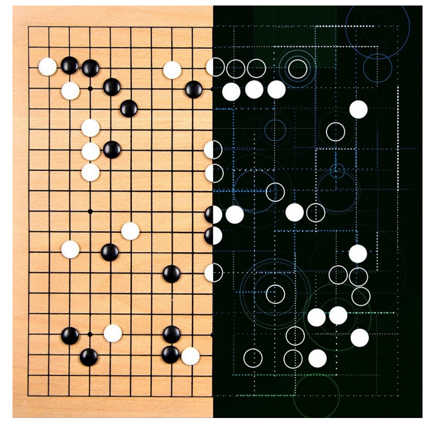
$$\mathrm{NEGap}(\pi) := \max_{i\in[m]}\left(\max_{\pi_i^\dagger} V_i^{\pi_i^\dagger,\pi_{-i}} - V_i^{\pi}\right) \le \varepsilon$$

i.e. each player plays the **best response** of all other player's policies.

🤔 What are natural learning goals in Markov Games?
(Generalizing "near-optimal policy" in MDPs)

Two-Player Zero-Sum Markov Games

Two-Player Zero-Sum Markov Games



Two-Player Zero-Sum MGs: $m = 2,\ r_1 \equiv 1 - r_2$

NE can be learned efficiently with polynomial time and samples:

[BT02, WHL17, JYM19, SMYY19, **B**J20, XCWY20, **B**JY20, ZKBY20, LY**B**J20, CZG21, JLY21, HLWZ21, LCWC22…]

Planning Algorithm

Nash Value Iteration (**Nash-VI**):

- Initialize $V^\star_{H+1}(s) \equiv 0$ for all $s \in \mathcal{S}$
- For $h = H, \ldots, 1$
  - For all (s, a_1, a_2):

    $Q^\star_h(s, a_1, a_2) = r_h(s, a_1, a_2) + (\mathbb{P}_h V^\star_{h+1})(s, a_1, a_2)$
  - For all s:

    $(\pi^\star_{1,h}(\cdot \,|\, s), \pi^\star_{2,h}(\cdot \,|\, s)) = \text{MatrixNash}(Q^\star_h(s, \cdot, \cdot))$

    $V^\star_h(s) = \langle \pi^\star_{1,h}(\cdot \,|\, s) \times \pi^\star_{2,h}(\cdot \,|\, s), Q^\star_h(s, \cdot, \cdot) \rangle$

Matrix Nash subroutine:

$$\text{MatrixNash}(Q) = \arg\Big(\max_{\pi_1 \in \Delta(\mathcal{A})} \min_{\pi_2 \in \Delta(\mathcal{B})} \langle \pi_1 \times \pi_2, Q \rangle \Big)$$

Nash-VI computes an exact NE (of a *known* game) in $\text{poly}(H, S, A_1, A_2)$ time.

🤔 Learn NE in *online setting* (only observe trajectories from playing)?

Optimistic Nash-VI

[Liu, Yu, **Bai**, Jin 2020]

- Initialize $\overline{Q}_{H+1}(s) \leftarrow H, \underline{Q}_{H+1}(s) \leftarrow 0$ for all $s \in \mathcal{S}$
- For episode $k = 1,\ldots,K$:
- For $h = H,\ldots,1$:
  - For all (s, a_1, a_2):

    $$\overline{Q}_h(s, a_1, a_2) = r_h(s, a_1, a_2) + (\hat{\mathbb{P}}_h \overline{V}_{h+1})(s, a_1, a_2) + \beta$$

    $$\underline{Q}_h(s, a_1, a_2) = r_h(s, a_1, a_2) + (\hat{\mathbb{P}}_h \underline{V}_{h+1})(s, a_1, a_2) - \beta$$

  - For all s:

    $$\pi_h(\cdot, \cdot \mid s) = \mathrm{MatrixCCE}(\overline{Q}_h(s, \cdot, \cdot), \underline{Q}_h(s, \cdot, \cdot))$$

    $$\overline{V}_h(s) = \langle \pi_h(\cdot, \cdot \mid s), \overline{Q}_h(s, \cdot, \cdot) \rangle$$

    $$\underline{V}_h(s) = \langle \pi_h(\cdot, \cdot \mid s), \underline{Q}_h(s, \cdot, \cdot) \rangle$$

- Play one episode using policy π, and update model estimate

Empirical model estimate

Optimistic bonus (Bernstein + model-based [DLWB18])

Coarse Correlated Equilibrium (CCE) subroutine [XCWY20]

Optimistic Nash-VI

[Liu, Yu, **Bai**, Jin 2020]

Theorem: Optimistic Nash-VI finds ε-NE within

$$K = \widetilde{O}\left(H^3 S A_1 A_2 / \varepsilon^2\right)$$

episodes of play.

✓ Learns NE in online setting with poly time & samples

✓ Natural extension of single-agent UCBVI algorithm [Azar et al. 2017]

✗ Compared with sample complexity lower bound $\Omega(H^3 S \max\{A_1, A_2\}/\varepsilon^2)$:

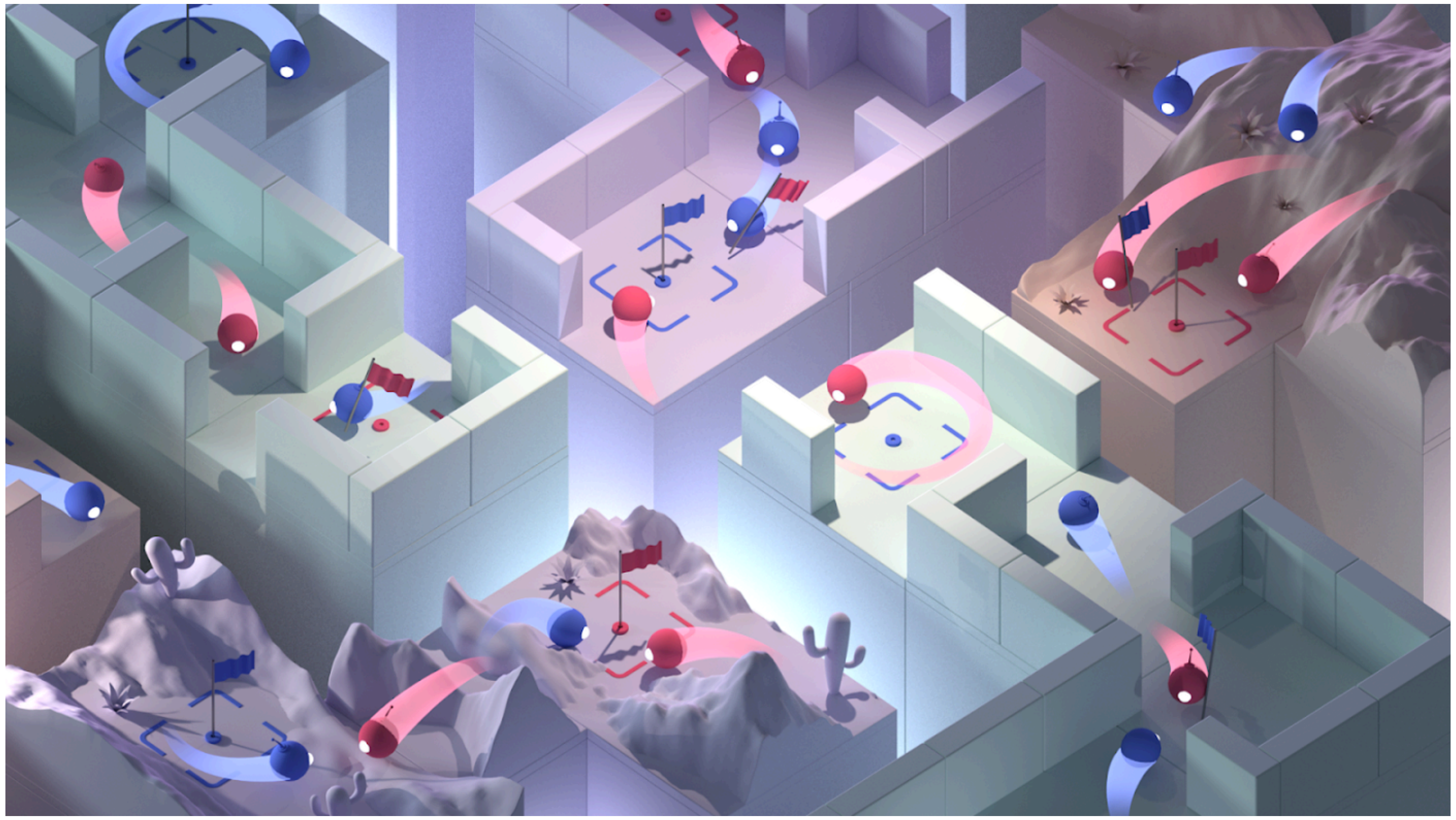
$$A_1 A_2 \quad \text{vs.} \quad \max\{A_1, A_2\}$$

😃 I'll show you another algorithm that

- Resolves this in the two-player zero-sum setting
- Provides new results in the multi-player general-sum setting

Multi-Player General-Sum Markov Games

Multi-Player General-Sum MGs



"Curse of Multiagents": |Joint action space| = **exp(# players)**

Learning NE in General-Sum MGs

Theorem [LYBJ20]: For general-sum MGs, Multi-Nash-VI finds ε-NE within

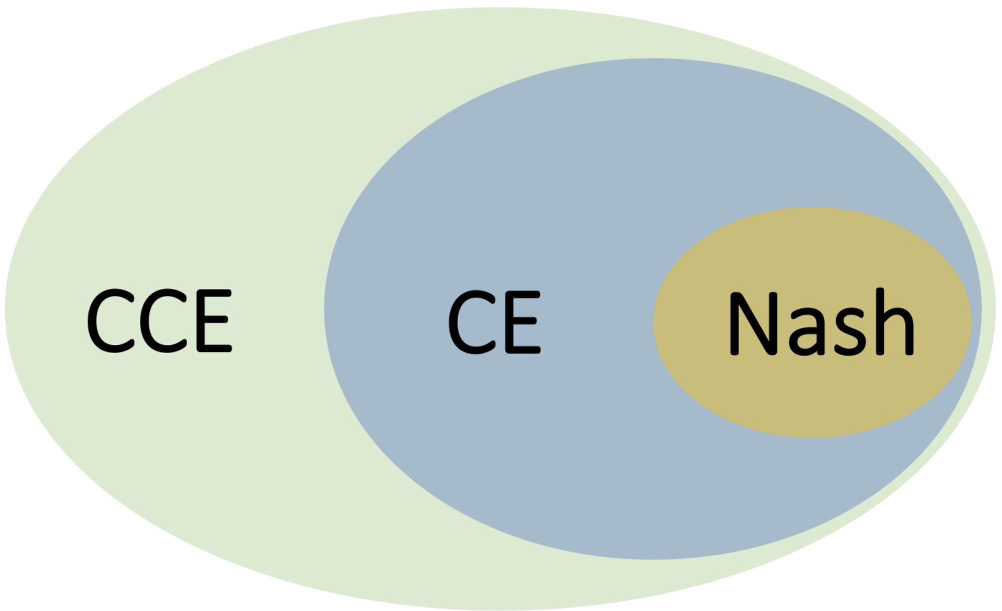
$$K = \widetilde{O}\left(H^4 S^2 \prod_{i \in [m]} A_i / \varepsilon^2\right)$$

episodes of play.

☹️ **Theorem** [Rubinstein 2016]: $\exp(\Omega(m))$ samples is unavoidable for learning NE even in multi-player general-sum **NFGs**.

Question: What equilibria can be learned with poly(m) samples?

Other Equilibria in Game Theory



Coarse Correlated Equilibrium (CCE):

No player gains by deviating from the correlated policy.

Correlated Equilibrium (CE):

No player gains by deviating from the correlated policy, even if the player observes her own sampled action .

Coarse Correlated Equilibria (CCE) in NFGs

Coarse Correlated Equilibrium (CCE): A *correlated policy* π is an ε-CCE if

$$\mathrm{CCEGap}(\pi) := \max_{i\in[m]}\left(\max_{\pi_i^\dagger} V_i^{\pi_i^\dagger,\pi_{-i}} - V_i^{\pi}\right) \le \varepsilon$$

No-regret to CCE: For NFGs, run **no-regret algorithm** for each player for T rounds, then $\widehat{\pi} := \mathrm{Unif}(\{\pi^t\}_{t=1}^T)$ satisfies

$$\mathrm{CCEGap}(\widehat{\pi}) = \max_{i\in[m]} \mathrm{Reg_i}(T)/T,$$

Corollary: Each player runs an **adversarial bandit algorithm** (e.g. EXP3),

$$\mathrm{CCEGap}(\widehat{\pi}) = \max_{i\in[m]} \mathrm{Reg_i}(T)/T \le \widetilde{O}\left(\sqrt{\max_{i\in[m]} A_i/T}\right)$$

Avoids curse of multiagent: Sample complexity depends on $\max_{i\in[m]} A_i$ only.

CCE in Markov Games

Coarse Correlated Equilibrium (CCE): A *correlated policy* π is an ε-CCE if

$$\mathrm{CCEGap}(\pi) := \max_{i \in [m]} \left(\max_{\pi_i^\dagger} V_i^{\pi_i^\dagger, \pi_{-i}} - V_i^{\pi} \right) \leq \varepsilon$$

Challenges for extending to Markov Games:

1. How to ensure **efficient exploration** (visit all relevant states)?
2. **No-regret in MGs** is intractable [Liu, Wang, Jin 2022]
   —what's the right goal / algorithm design?
3. (Side quest) **Decentralized algorithm**?

😀 Were addressed in two-player zero-sum MGs:

Nash V-Learning algorithm [**Bai**, Jin, Yu 2020]

Nash V-Learning (max-player) **for zero-sum MGs**

1. Maintain optimistic V values with incremental update (≈ Q-Learning)

$$\overline{V}_h(s_h) \leftarrow (1-\alpha_t)\overline{V}_h(s_h) + \alpha_t(r_h + \overline{V}_{h+1}(s_{h+1}) + \text{bonus}(t))$$

when s_h is visited for t-th time.

Ensures exploration

2. Update policy by adversarial bandit subroutine at (h, s_h):

$$\mu_h(\cdot \mid s_h) \leftarrow \text{Adv\_Bandit\_Update}(a_h, \frac{H - r_h - \overline{V}_{h+1}(s_{h+1})}{H})$$

(e.g. weighted anytime FTRL).

Achieves "per-state" regrets

3. Play an episode with policy μ, observe transitions, rewards
4. After K episodes, output *certified policy* $\hat{\mu}$

Nash V-Learning (max-player) **for zero-sum MGs**

1. Maintain optimistic V values with incremental update (≈ Q-Learning)

$$\overline{V}_h(s_h) \leftarrow (1-\alpha_t)\overline{V}_h(s_h) + \alpha_t(r_h + \overline{V}_{h+1}(s_{h+1}) + \mathrm{bonus}(t))$$

   when s_h is visited for t-th time.

2. Update policy by adversarial bandit subroutine at (h, s_h):

$$\mu_h(\cdot \mid s_h) \leftarrow \mathrm{Adv\_Bandit\_Update}\left(a_h, \frac{H - r_h - \overline{V}_{h+1}(s_{h+1})}{H}\right)$$

   (e.g. weighted anytime FTRL).

3. Play an episode with policy μ, observe transitions, rewards
4. After K episodes, output *certified policy* $\hat{\mu}$

Theorem [**Bai**, Jin, Yu 2020]: Nash V-Learning finds ε-NE within

$$K = \widetilde{O}\left(H^5 S \max\{A_1, A_2\}/\varepsilon^2\right)$$

episodes of play in zero-sum MGs.

CCE-V-Learning (i-th player) **for general-sum MGs**

1. Maintain optimistic V values with incremental update
$$\overline{V}_{i,h}(s_h) \leftarrow (1-\alpha_t)\overline{V}_{i,h}(s_h) + \alpha_t(r_{i,h} + \overline{V}_{i,h+1}(s_{h+1}) + \mathrm{bonus}(t))$$
when s_h is visited for t-th time.

2. Update policy by adversarial bandit subroutine at (h, s_h):
$$\pi_{i,h}(\cdot \mid s_h) \leftarrow \mathrm{Adv\_Bandit\_Update}(a_{i,h}, \frac{H - r_{i,h} - \overline{V}_{i,h+1}(s_{h+1})}{H})$$
(e.g. weighted anytime FTRL).

3. Play an episode with policy π_i, observe transitions, rewards
4. After K episodes, output *certified correlated policy* $\widehat{\pi}$

Theorem [Song, Mei, **Bai** 2021]: CCE-V-Learning finds ε-CCE within
$$K = \widetilde{O}\left(H^5 S(\max_{i\in[m]} A_i)/\varepsilon^2\right)$$
episodes of play in general-sum MGs.

CCE-V-Learning (i-th player) **for general-sum MGs**

Theorem [Song, Mei, **Bai** 2021]: CCE-V-Learning finds ε-CCE within

$$K = \widetilde{O}\left(H^5 S(\max_{i\in[m]} A_i)/\varepsilon^2\right)$$

episodes of play in general-sum MGs.

✓ Avoids curse-of-multiagent: $\mathrm{poly}(H, S, \max_{i\in[m]} A_i, 1/\varepsilon^2)$ samples

✓ Learns in online/exploration setting

✓ Decentralized algorithm

✗ Output policy is non-Markov (history-dependent)

🤔 Markov CCE can be learned by VI / “stage-wise” algorithms:

$\widetilde{O}(\prod_{i\in[m]} A_i/\varepsilon^2)$ sample complexity [Liu, Yu, **Bai**, Jin 2020]

$\widetilde{O}(\max_{i\in[m]} A_i/\varepsilon^3)$ by recent work of [Daskalakis, Golowich, Zhang 2022]

Extension to CE

Algorithm (CE-V-Learning, i-th player):

2'. Update policy by adversarial bandit subroutine at (h, s_h):

$$\pi_{i,h}(\cdot \mid s_h) \leftarrow \mathrm{Adv\_Bandit\_Update}\left(a_{i,h}, \frac{H - r_{i,h} - \overline{V}_{i,h+1}(s_{h+1})}{H}\right)$$

that minimizes weighted swap regret (e.g. mixed-expert FTRL [Ito 2020])

Theorem [Song, Mei, **Bai** 2021]: CE-V-Learning finds ε-CE within

$$K = \widetilde{O}\left(H^6 S (\max_{i\in[m]} A_i^2)/\varepsilon^2\right)$$

episodes of play in general-sum MGs.

Literature note

1. *When Can We Learn General-Sum Markov Games with A Large Number of Players Sample-Efficiently?*
   Ziang Song, Song Mei, Yu Bai. arXiv:2110.04184.
   → Contains CE/CCE results.

2. *V-Learning—A Simple, Efficient, Decentralized Algorithm for Multiagent RL.*
   Chi Jin, Qinghua Liu, Yuanhao Wang, Tiancheng Yu. arXiv:2110.14555.
   → Contains CE/CCE results, with H-better rate for CE (different swap-regret alg.)

3. *Provably Efficient Reinforcement Learning in Decentralized General-Sum Markov Games.*
   Weichao Mao, Tamer Başar. arXiv:2110.05682.
   → Contains CCE results.

All 3 papers are based on the V-Learning algorithm proposed in

Near-Optimal Reinforcement Learning with Self-Play.
Yu Bai, Chi Jin, Tiancheng Yu. NeurIPS 2020.
(NE for two-player zero-sum Markov Games)

Faster Convergence via Optimistic Algorithms

Learning NFGs under full-information feedback

Hedge (FTRL) Algorithm:

For $t = 1, \ldots, T$:

- Receive utility vector based on opponents' strategies:

$$u_i^t(a) = r_i(a, \pi_{-i}^t)$$

- Update strategy by exponential weights:

$$\pi_i^{t+1}(a) \propto_a \pi_i^t(a) \cdot \exp(\eta u_i^t(a))$$

Hedge achieves $O(\sqrt{T})$ regret against **any** seq. of opponents (e.g. [CBL06])

Corollary: Let all players play Hedge against each other,

- Learns CCE in NFGs with $O(T^{-1/2})$ convergence rate
- Learns NE in two-player zero-sum NFGs with $O(T^{-1/2})$ convergence rate

Issues with Hedge approach

Hedge regret bound works for any adversarial opponent

Analysis does not use that opponents are also playing Hedge

🤔 Can we get faster convergence to NE/CCE if we use the fact that everyone is playing the same no-regret algorithm?

Optimistic Hedge / OFTRL

Optimistic Hedge (OFTRL) Algorithm:

- Update strategy by exponential weights over lookahead adjusted utility vector

$$\pi_i^{t+1}(a) \propto_a \pi_i^t(a) \cdot \exp(\eta(2u_i^t(a) - u_i^{t-1}(a)))$$

Intuition: When u_i^t changes slowly in t,

$$2u_i^t - u_i^{t-1} = u_i^t + (u_i^t - u_i^{t-1}) \approx u_i^t + (u_i^{t+1} - u_i^t) = u_i^{t+1}$$

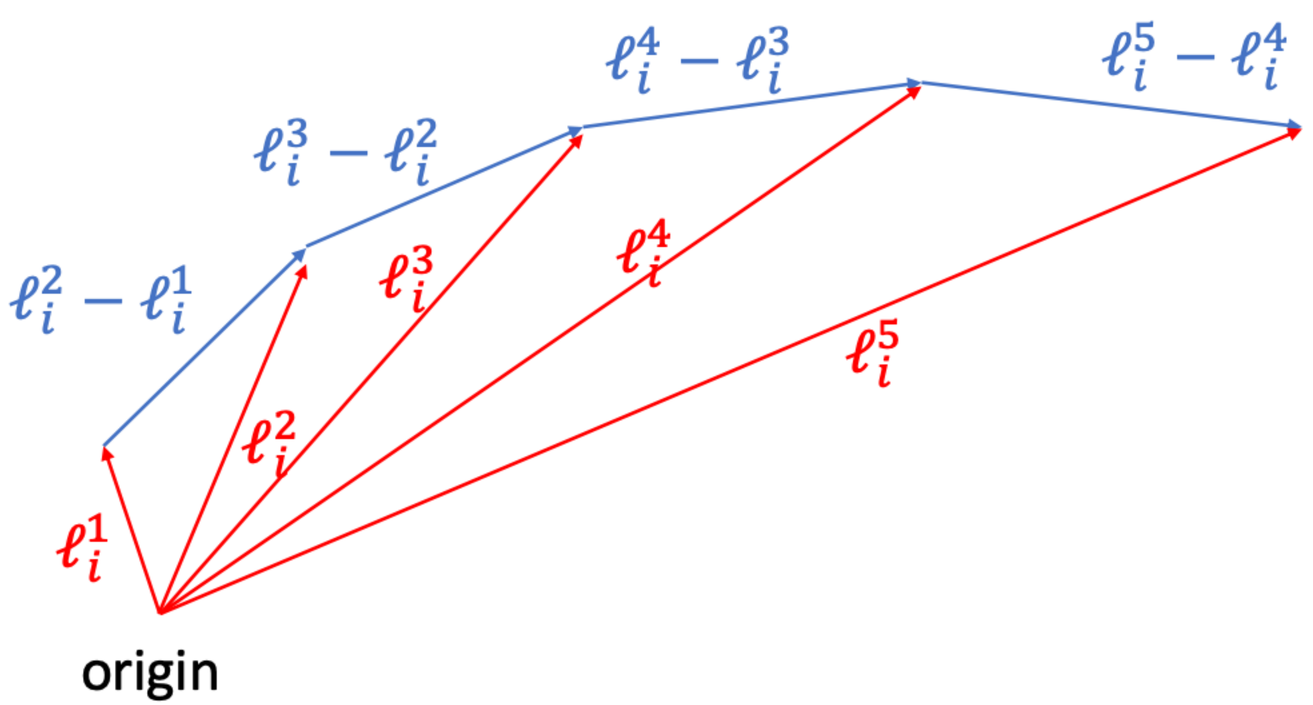


Image source:
Min-Max Optimization (Simons Institute), Costis Daskalakis,, 2022.

Regret Bounds of Optimistic Algorithms in Games

Table 1: Overview of prior work on fast rates for learning in games. m denotes the number of players, and n denotes the number of actions per player (assumed to be the same for all players). For Optimistic Hedge, the adversarial regret bounds in the right-hand column are obtained via a choice of adaptive step-sizes. The $\tilde{O}(\cdot)$ notation hides factors that are polynomial in $\log T$.

Algorithm	Setting	Regret in games	Adversarial regret
Hedge (& many other algs.)	multi-player, general-sum	$O(\sqrt{T \log n})$ [CBL06]	$O(\sqrt{T \log n})$ [CBL06]
Excessive Gap Technique	2-player, 0-sum	$O(\log n (\log T + \log^{3/2} n))$ [DDK11]	$O(\sqrt{T \log n})$ [DDK11]
DS-OptMD, OptDA	2-player, 0-sum	$\log^{O(1)}(n)$ [HAM21]	$\sqrt{T \log^{O(1)}(n)}$ [HAM21]
Optimistic Hedge	multi-player, general-sum	$O(\log n \cdot \sqrt{m} \cdot T^{1/4})$ [RS13b, SALS15]	$\tilde{O}(\sqrt{T \log n})$ [RS13b, SALS15]
Optimistic Hedge	2-player, general-sum	$O(\log^{5/6} n \cdot T^{1/6})$ [CP20]	$\tilde{O}(\sqrt{T \log n})$
Optimistic Hedge	multi-player, general-sum	$O(\log n \cdot m \cdot \log^4 T)$ **(Theorem 3.1)**	$\tilde{O}(\sqrt{T \log n})$ **(Corollary D.1)**

Breakthrough paper:

- **Near-Optimal No-Regret Learning in General Games.**
  Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich.
  In NeurIPS 2021 **(Oral presentation)**. [conf]

Near-optimal no-regret learning in general games
C Daskalakis, M Fishelson… - Advances in Neural …, 2021 - proceedings.neurips.cc
Abstract We show that Optimistic Hedge--a common variant of multiplicative-weights-updates with recency bias--attains ${\rm poly}(\log T) $ regret in multi-player general-sum games. In particular, when every player of the game uses Optimistic Hedge to iteratively update her action in response to the history of play so far, then after $ T $ rounds of interaction, each player experiences total regret that is ${\rm poly}(\log T) $. Our bound improves, exponentially, the $ O (T^{1/2}) $ regret attainable by standard no-regret learners …

Faster Convergence to NE/CCE in NFGs

[Daskalakis, Fishelson, Golowich 2021]

OFTRL achieves $O(\log^4 T) = \widetilde{O}(1)$ regret when played by everyone in a game.

Corollary: Let all players play OFTRL against each other,

- Learn CCE with $\widetilde{O}(T^{-1})$ convergence rate
- Learn NE in two-player zero-sum games with $\widetilde{O}(T^{-1})$ convergence rate*

* Also well-established e.g. [RS13b] by a more direct analysis for zero-sum case

Question: Extend to Markov Games?

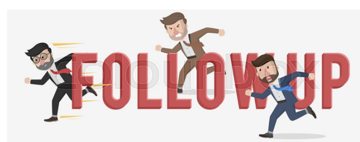
Faster Convergence to NE/CCE in Markov Games

[Zhang*, Liu*, Wang, Xiong, Li, **Bai** NeurIPS 2022]

Theorem: We obtain faster convergence results for MGs:

- $\widetilde{O}(T^{\{-5/6,-1\}})$ for learning NE in two-player zero-sum MGs
- $\widetilde{O}(T^{-3/4})$ for learning CCE in multi-player general-sum MGs

Algorithm is natural: OFTRL + smooth value updates

Immediate 

s:

$O(T^{-1})$ Convergence of Optimistic-Follow-the-Regularized-Leader in Two-Player Zero-Sum Markov Games

Yuepeng Yang*   Cong Ma*

September 27, 2022

Abstract

We prove that optimistic-follow-the-regularized-leader (OFTRL), together with smooth value updates, finds an $O(T^{-1})$-approximate Nash equilibrium in T iterations for two-player zero-sum Markov games with full information. This improves the $\tilde{O}(T^{-5/6})$ convergence rate recently shown in the paper [ZLW+22]. The refined analysis hinges on two essential ingredients. First, the sum of the regrets of the two players, though not necessarily non-negative as in normal-form games, is approximately non-negative in Markov games. This property allows us to bound the second-order path lengths of the learning dynamics. Second, we prove a tighter algebraic inequality regarding the weights deployed by OFTRL that shaves an extra $\log T$ factor. This crucial improvement enables the inductive analysis that leads to the final $O(T^{-1})$ rate.

Faster Last-iterate Convergence of Policy Optimization in Zero-Sum Markov Games

Shicong Cen[1*]   Yuejie Chi[1†]   Simon S. Du[2,3‡]   Lin Xiao[3§]

[1]Carnegie Mellon University
[2]University of Washington
[3]Meta AI Research

October 5, 2022

Regret Minimization and Convergence to Equilibria in General-sum Markov Games

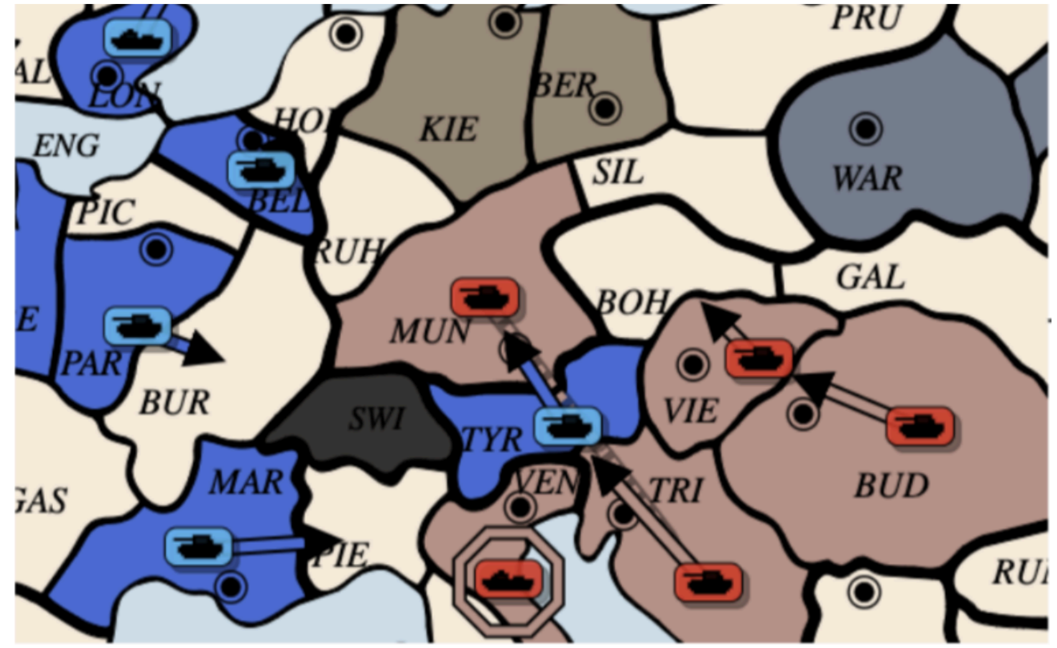
Liad Erez[1,*]   Tal Lancewicki[1,*]   Uri Sherman[1,*]   Tomer Koren[1,2]

Yishay Mansour[1,2]

August 9, 2022

Advanced Topics

Imperfect Information



Imperfect Information / Partial Observability:

Players can only observe *partial information* about the true underlying game

Recent advances in Poker [Moravcik et al. 2017, Brown & Sandholm 2018, 2019], Bridge [Tian et al. 2020], Diplomacy [Bakhtin et al. 2021], …

Formulation: Imperfect-Information Extensive-Form Games (EFGs)

Image source (right):
No-Press Diplomacy from Scratch, Bakhtin et al. 2021.

Learning EFGs from bandit feedback

Algorithm	Equilibrium	Sample Complexity
Farina et al. [2021]	CCE	$\widetilde{O}(X^4 A^3/\varepsilon^2)$
Kozuno et al. [2021]	CCE	$\widetilde{O}(X^2 A/\varepsilon^2)$
Bai, Jin, Mei, Yu [2022]	CCE	$\widetilde{O}(XA/\varepsilon^2)$
Song, Mei, **Bai** [2022]	K-EFCE*	$\widetilde{O}(XA^{K+1}/\varepsilon^2)$
Bai, Jin, Mei, Song, Yu [2022]	EFCE	$\widetilde{O}(XA/\varepsilon^2)$

X: number of information sets; A: number of actions

* Newly defined equilibrium, {K-EFCE}⊂{1-EFCE}⊂{EFCE}

Building on two main EFG algorithms (full-information setting):

- Online Mirror Descent [Hoda et al. 2010, Kroer et al. 2015]
- Counterfactual Regret Minimization [Zinkevich et al. 2007, Celli et al. 2020]

Heavily rely on tree structure of EFGs, which **do not hold** in general POMGs.

Dominance and Rationalizability

CCE (and approximate CE) can be supported entirely on dominated actions !

[Viossat & Zapechelnyuk 2013]

	b_1	b_2	b_3	b_4
a_1	1, 1	1, 1	1, 0	5, 1
a_2	1, 1	1, 1	5, 0	1, 0
a_3	0, 1	0, 5	4, 4	0, 0
a_4	0, 5	0, 1	0, 0	4, 4

Learning Rationalizable Equilibria

[Wang, Kong, **Bai**, Jin 2022]

Def: An action is rationalizable if it survives Iterative Dominance Elimination .

[Bernheim 1984; Pearce 1984]

We design the first algorithms for efficiently learning **ε-CE/CCE supported on Δ-rationalizable actions** in multi-player NFGs from bandit feedback.

(Related: Wu et al. [2021] find **any** rationalizable strategy, not nece. CE/CCE)

Task		Sample Complexity
Find *all* rationalizable actions (Proposition 3)		$\Omega(A^{N-1})$
Find *one* rationalizable action profile (Theorem 4)		$\widetilde{O}\left(\frac{LNA}{\Delta^2}\right)$
Learn rationalizable equilibria	ϵ-CCE (Theorem 7)	$\widetilde{O}\left(\frac{LNA}{\Delta^2} + \frac{NA}{\epsilon^2}\right)$
	ϵ-CE (Theorem 12)	$\widetilde{O}\left(\frac{LNA}{\Delta^2} + \frac{NA^2}{\min\{\epsilon^2, \Delta^2\}}\right)$

Table 1: Summary of main results. Here N is the number of players, A is the number of actions per player, $L < NA$ is the minimum elimination length and Δ is the error we allow for rationalizability.

Conclusion

My Excitement About MARL/Games:

1. Single-agent RL results can be (non-trivially) extended to MARL/games
   - e.g. Learning NE/CE/CCE in Markov Games

2. Games pose interesting questions to {online learning, bandits, RL…}
   - e.g. Faster no-regret learning when everyone runs a no-regret algorithm

3. Games admit unique questions that are potentially rich for ML theory:
   - e.g. Rationalizability

Open Questions

- **Function approximation**
  - "Reduce" to centralized single-agent problem
  - Decentralized / independent function approximation?
- **Imperfect information / partial observability**
  - EFGs
  - General Partially Observable Markov Games
- **Solution concepts beyond NE/CE/CCE**
  - General Φ-equilibria
  - Stackelberg Equilibria
  - Economics connections (e.g. rationalizability, contract theory)
- **Other types of games**
  - Markov potential games
  - Congestion games

Thank you!

Backup Slides

Certified Policies

Algorithm 2 Certified correlated policy $\widehat{\pi}$ for general-sum MGs

1: Sample $k \leftarrow \text{Uniform}([K])$.
2: **for** step $h = 1, \ldots, H$ **do**
3: Observe s_h, and set $t \leftarrow N_h^k(s_h)$ (the value of $N_h(s_h)$ at the beginning of the k'th episode).
4: Sample $l \in [t]$ with $\mathbb{P}(l = j) = \alpha_t^j$ (c.f. Eq. (3)).
5: Update $k \leftarrow k_h^l(s_h)$ (the episode at the end of which the state s_h is observed exactly l times).
6: Jointly take action $(a_{h,1}, a_{h,2}, \ldots, a_{h,m}) \sim \prod_{i=1}^{m} \mu_{h,i}^k(\cdot|s_h)$, where $\mu_{h,i}^k(\cdot|s_h)$ is the policy $\mu_{h,i}(\cdot|s_h)$ at the beginning of the k'th episode.