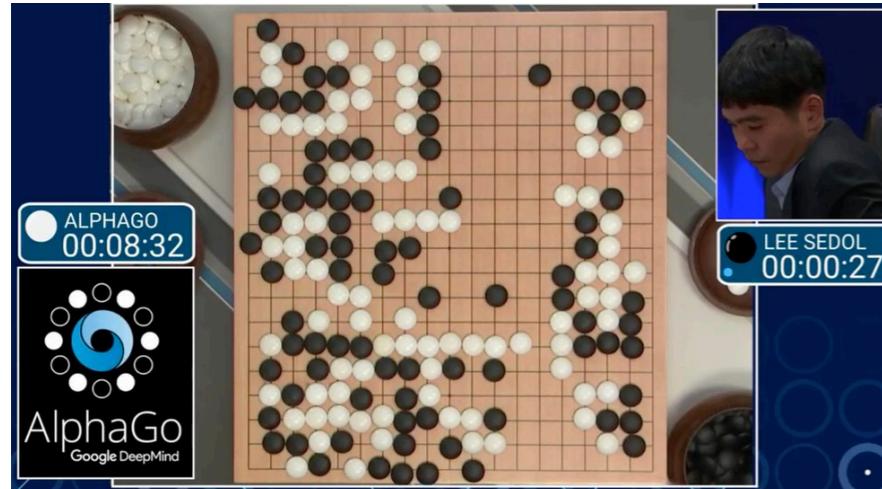


# **Recent Progresses on the Theory of Multi-Agent Reinforcement Learning and Games**

**Yu Bai**  
Salesforce Research

Blog post: [https://yubai.org/blog/marl\\_theory.html](https://yubai.org/blog/marl_theory.html)

# Multi-Agent Reinforcement Learning



AlphaGo



Poker



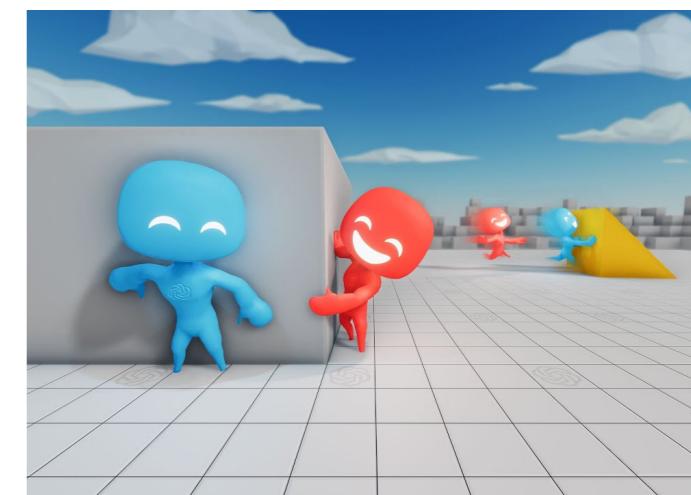
AI Economist



Starcraft

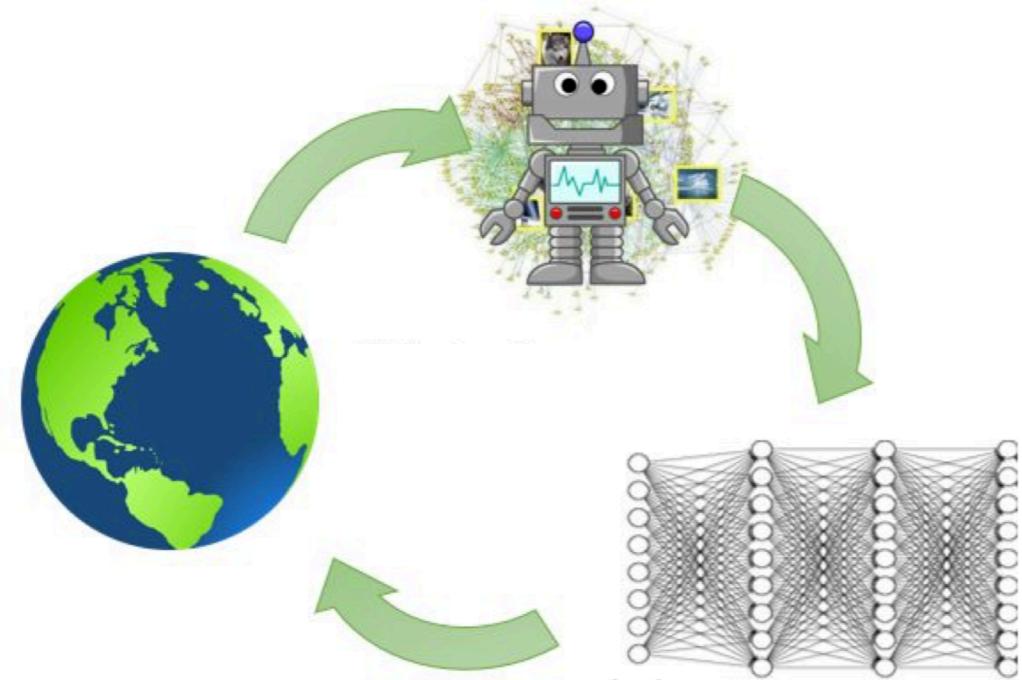


Diplomacy

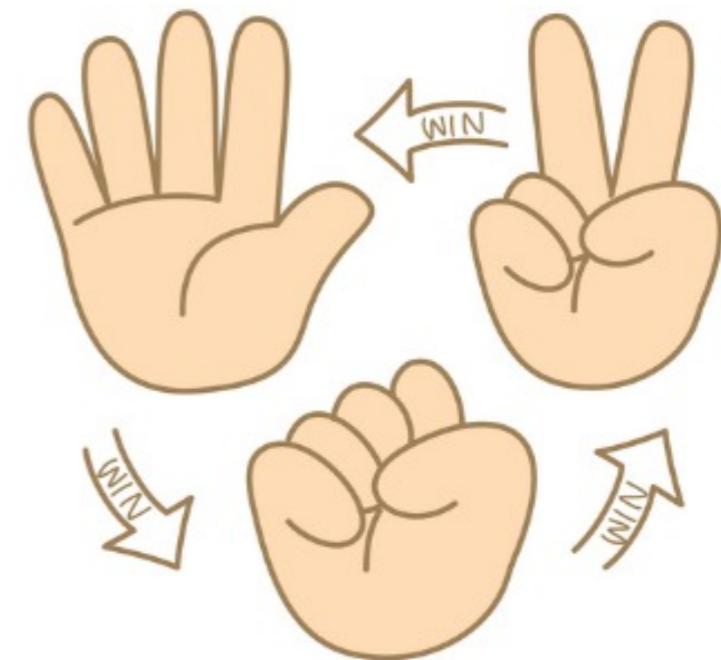


Hide and Seek

# Multi-Agent Reinforcement Learning



sequential decisions



multi-agent

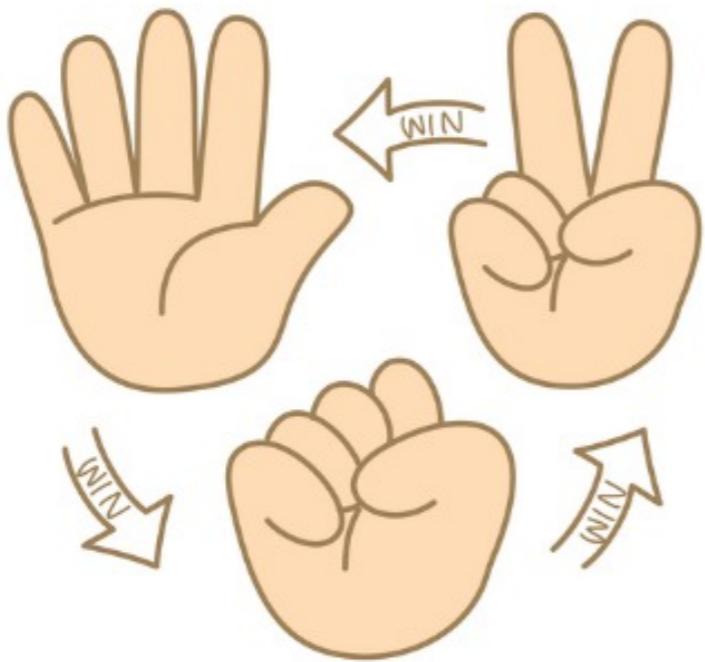
A relatively **new** field, with **unique challenges** and opportunities for both **theory**/empirical research.

# 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

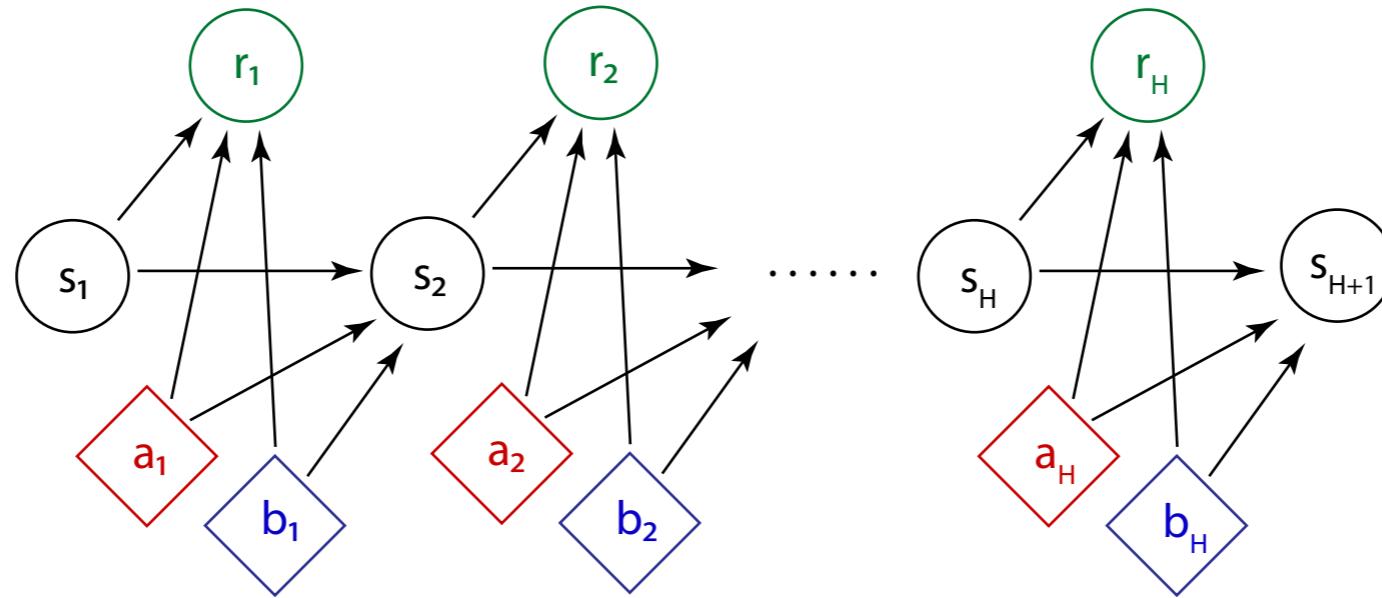
Multi-player Normal-Form Games (NFGs):

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

# Markov Games (MGs)

[Shapley 1953]

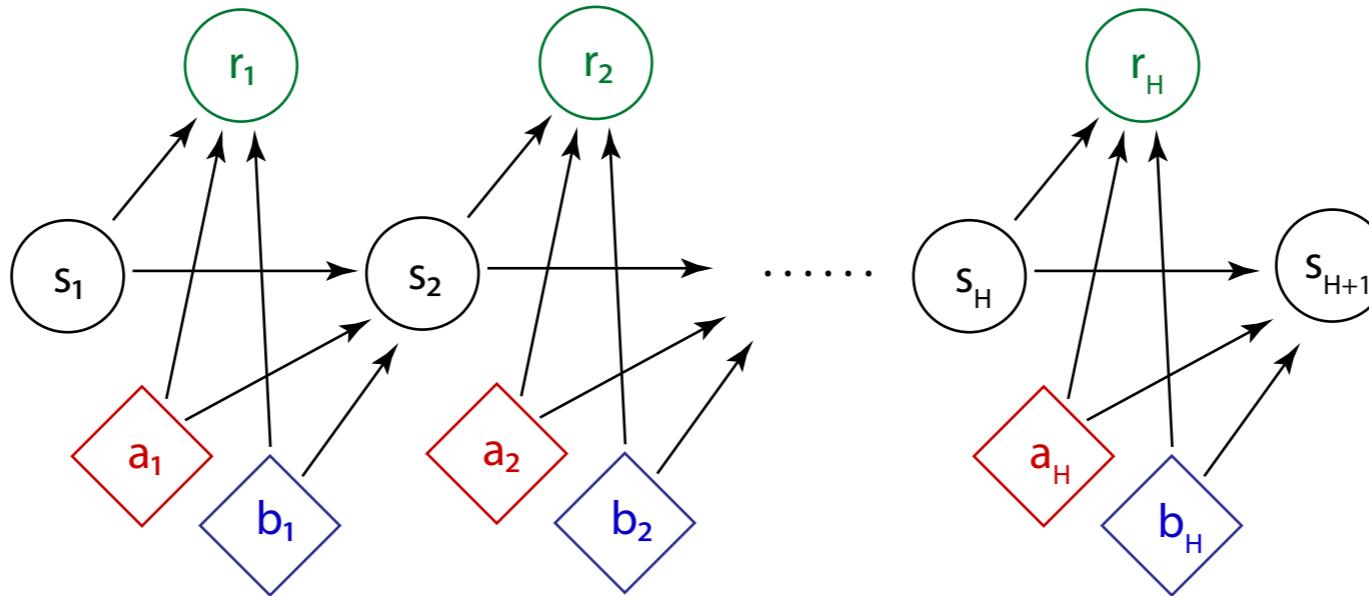
(also known as **Stochastic Games**)



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:  $r_{i,h}(s_h, a_{1,h}, \dots, a_{m,h})$  (for  $i$ -th player)
- Transition:  $(s_h, a_{1,h}, \dots, a_{m,h}) \rightarrow s_{h+1}$

# Policies, Values, Equilibria



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

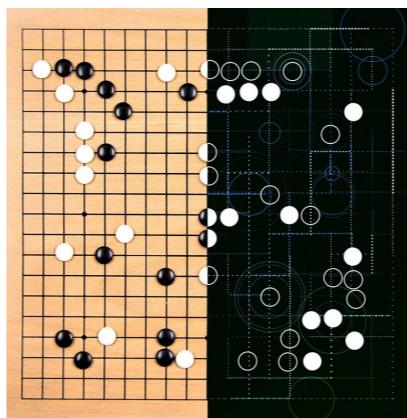
**Nash Equilibrium (NE):** A product policy  $\pi = \{\pi_i\}_{i \in \mathcal{I}}$  is an  $\varepsilon$ -NE if

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

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

# 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, **BJ**20, XCWY20, **BJY**20, ZKBY20, LY**BJ**20, CZG21, JLY21, HLWZ21, LCWC22...]



# Planning Algorithm

## Nash Value Iteration (Nash-VI):

- Initialize  $V_{H+1}^*(s) \equiv 0$  for all  $s \in \mathcal{S}$

- For  $h = H, \dots, 1$

- For all  $(s, a_1, a_2)$ :

$$Q_h^*(s, a_1, a_2) = r_h(s, a_1, a_2) + (\mathbb{P}_h V_{h+1}^*)(s, a_1, a_2)$$

- For all  $s$ :

$$(\pi_{1,h}^*(\cdot | s), \pi_{2,h}^*(\cdot | s)) = \text{MatrixNash}(Q_h^*(s, \cdot, \cdot))$$

$$V_h^*(s) = \langle \pi_{1,h}^*(\cdot | s) \times \pi_{2,h}^*(\cdot | s), Q_h^*(s, \cdot, \cdot) \rangle$$

Matrix Nash subroutine:

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

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  $\bar{Q}_{H+1}(s) \leftarrow H, \underline{Q}_{H+1}(s) \leftarrow 0$  for all  $s \in \mathcal{S}$

- For episode  $k = 1, \dots, K$ :

- For  $h = H, \dots, 1$ :

- For all  $(s, a_1, a_2)$ :

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

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

- For all  $s$ :

$$\pi_h(\cdot, \cdot | s) = \text{MatrixCCE}(\bar{Q}_h(s, \cdot, \cdot), \underline{Q}_h(s, \cdot, \cdot))$$

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

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

Empirical model estimate

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

Coarse Correlated Equilibrium (CCE) subroutine [XCWY20]

- Play one episode using policy  $\pi$ , and update model estimate

# Optimistic Nash-VI

[Liu, Yu, Bai, Jin 2020]

**Theorem:** Optimistic Nash-VI finds  $\varepsilon$ -NE within

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

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  $\Omega(H^3 S \max\{A_1, A_2\} / \varepsilon^2)$ :  
 $\overline{A_1 A_2}$  vs.  $\overline{\max\{A_1, A_2\}}$
- 😊 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”:**  $|\text{Joint action space}| = \exp(\# \text{ players})$

# Learning NE in General-Sum MGs

**Theorem [LYBJ20]:** For general-sum MGs, Multi-Nash-VI finds  $\varepsilon$ -NE within

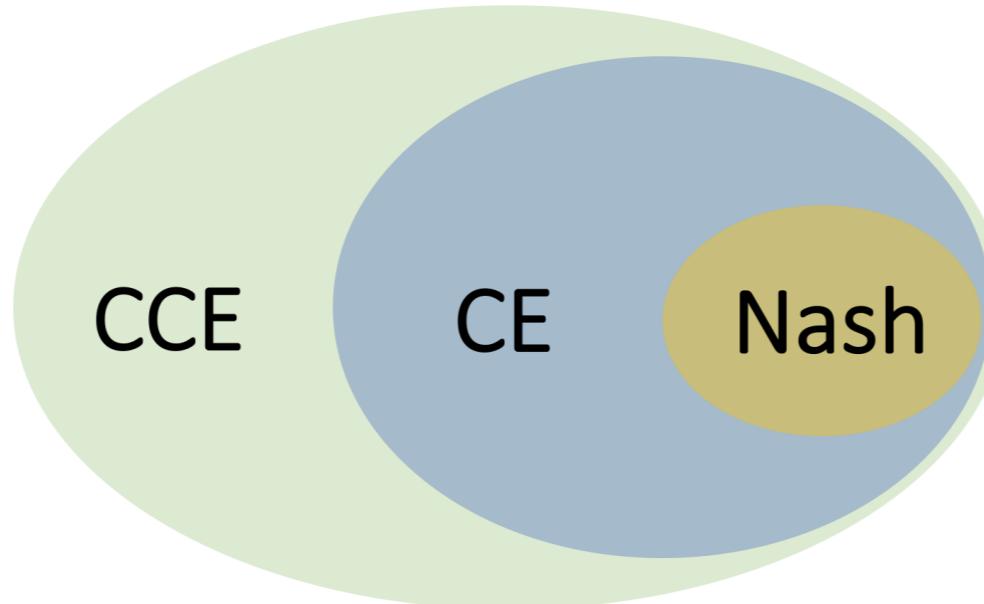
$$K = \widetilde{O}\left(H^4 S^2 \prod_{i \in [m]} A_i / \varepsilon^2\right)$$

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  $\text{poly}(m)$  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  $\pi$  is an  $\varepsilon$ -CCE if

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

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

$$\text{CCEGap}(\hat{\pi}) = \max_{i \in [m]} \text{Reg}_i(T)/T,$$

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

$$\text{CCEGap}(\hat{\pi}) = \max_{i \in [m]} \text{Reg}_i(T)/T \leq \widetilde{O} \left( \sqrt{\max_{i \in [m]} A_i / T} \right)$$

**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  $\pi$  is an  $\varepsilon$ -CCE if

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

**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 ( $\approx$  Q-Learning)

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

when  $s_h$  is visited for  $t$ -th time.

Ensures exploration

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

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

(e.g. weighted anytime FTRL).

Achieves “per-state” regrets

3. Play an episode with policy  $\mu$ , 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 ( $\approx$  Q-Learning)

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

when  $s_h$  is visited for  $t$ -th time.

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

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

(e.g. weighted anytime FTRL).

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

**Theorem** [Bai, Jin, Yu 2020]: Nash V-Learning finds  $\varepsilon$ -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

$$\bar{V}_{i,h}(s_h) \leftarrow (1 - \alpha_t) \bar{V}_{i,h}(s_h) + \alpha_t (r_{i,h} + \bar{V}_{i,h+1}(s_{h+1}) + \text{bonus}(t))$$

when  $s_h$  is visited for  $t$ -th time.

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

$$\pi_{i,h}(\cdot | s_h) \leftarrow \text{Adv\_Bandit\_Update}(a_{i,h}, \frac{H - r_{i,h} - \bar{V}_{i,h+1}(s_{h+1})}{H})$$

(e.g. weighted anytime FTRL).

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

**Theorem** [Song, Mei, Bai 2021]: CCE-V-Learning finds  $\varepsilon$ -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  $\varepsilon$ -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**:  $\text{poly}(H, S, \max_{i \in [m]} A_i, 1/\varepsilon^2)$  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  $(h, s_h)$ :

$$\pi_{i,h}(\cdot | s_h) \leftarrow \text{Adv\_Bandit\_Update}(a_{i,h}, \frac{H - r_{i,h} - \bar{V}_{i,h+1}(s_{h+1})}{H})$$

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

**Theorem** [Song, Mei, Bai 2021]: CE-V-Learning finds  $\varepsilon$ -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  $t = 1, \dots, T$ :

- 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  $t$ ,

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

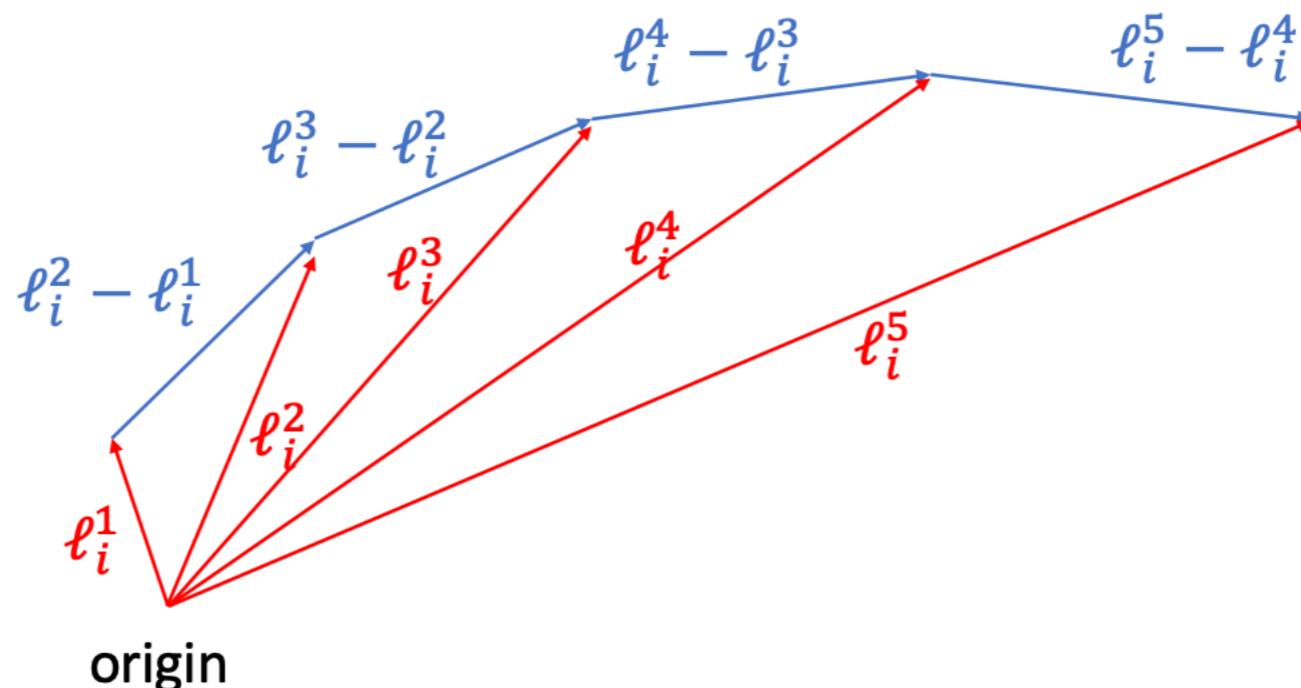


Image source:

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  $\mathcal{O}(\log T)$  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  $\mathcal{O}(\log T)$ . Our bound improves, exponentially, the  $\mathcal{O}(T^{1/2})$  regret attainable by standard no-regret learners ...

☆ Save ⚡ Cite Cited by 29 Related articles All 4 versions ☰

# 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:

- $\tilde{O}(T^{\{-5/6, -1\}})$  for learning NE in two-player zero-sum MGs
- $\tilde{O}(T^{-3/4})$  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<sup>+</sup>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 tightened 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<sup>1\*</sup> Yuejie Chi<sup>1†</sup> Simon S. Du<sup>2,3‡</sup> Lin Xiao<sup>3§</sup>

<sup>1</sup>Carnegie Mellon University

<sup>2</sup>University of Washington

<sup>3</sup>Meta AI Research

October 5, 2022

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

Liad Erez<sup>1,\*</sup>

Tal Lancewicki<sup>1,\*</sup>

Uri Sherman<sup>1,\*</sup>

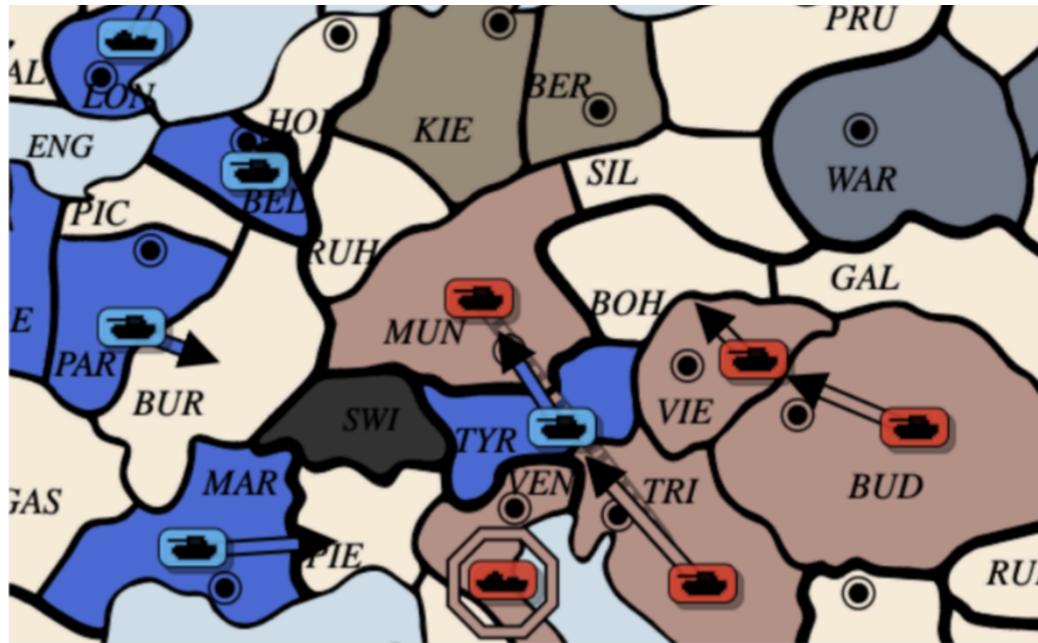
Tomer Koren<sup>1,2</sup>

Yishay Mansour<sup>1,2</sup>

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)

# 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	$\widetilde{O}(X^2 A / \varepsilon^2)$
<b>Bai, Jin, Mei, Yu [2022]</b>	CCE	$\widetilde{O}(XA / \varepsilon^2)$
Song, Mei, <b>Bai</b> [2022]	K-EFCE*	$\widetilde{O}(XA^{K+1} / \varepsilon^2)$
<b>Bai, Jin, Mei, Song, Yu [2022]</b>	EFCE	$\widetilde{O}(XA / \varepsilon^2)$

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

\* Newly defined equilibrium,  $\{\text{K-EFCE}\} \subset \{\text{1-EFCE}\} \subset \{\text{EFCE}\}$

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  **$\varepsilon$ -CE/CCE supported on  $\Delta$ -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 <i>all</i> rationalizable actions (Proposition 3)	$\Omega(A^{N-1})$
Find <i>one</i> rationalizable action profile (Theorem 4)	$\tilde{O}\left(\frac{LNA}{\Delta^2}\right)$
Learn rationalizable equilibria	$\tilde{O}\left(\frac{LNA}{\Delta^2} + \frac{NA}{\epsilon^2}\right)$
	$\tilde{O}\left(\frac{LNA}{\Delta^2} + \frac{NA^2}{\min\{\epsilon^2, \Delta^2\}}\right)$

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  $\Delta$  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  $\Phi$ -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  $\hat{\pi}$  for general-sum MGs

---

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