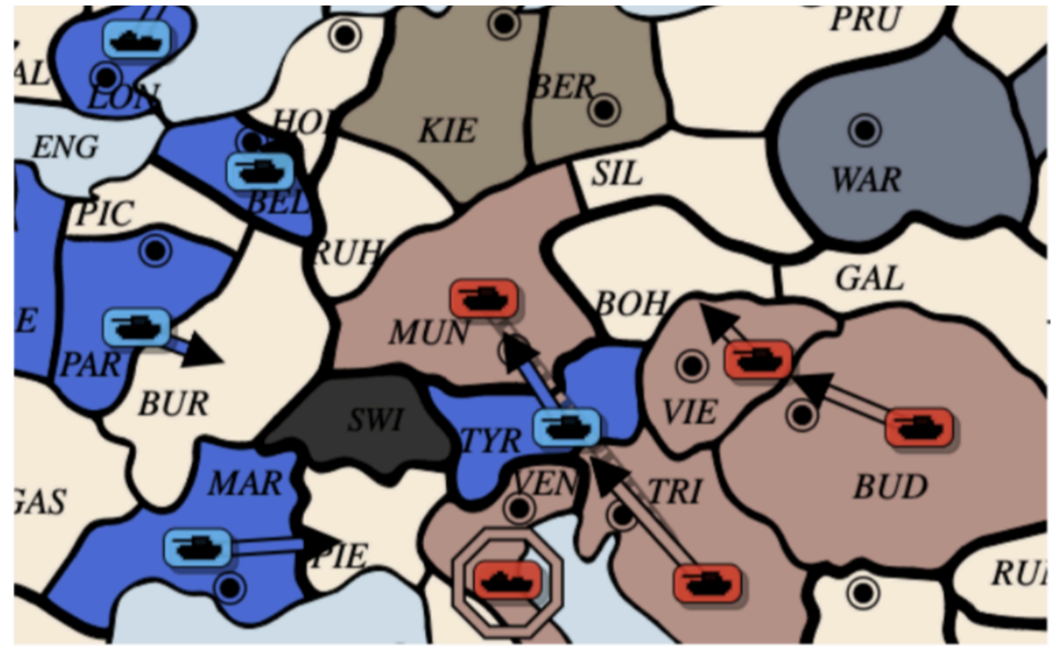
Near-Optimal Learning of Extensive-Form Games with Imperfect Information

Tiancheng Yu
MIT

Joint work with Yu Bai (Salesforce Research),
Chi Jin (Princeton) and Song Mei (UC Berkeley)

Multi-Agent RL / Games with Imperfect Information



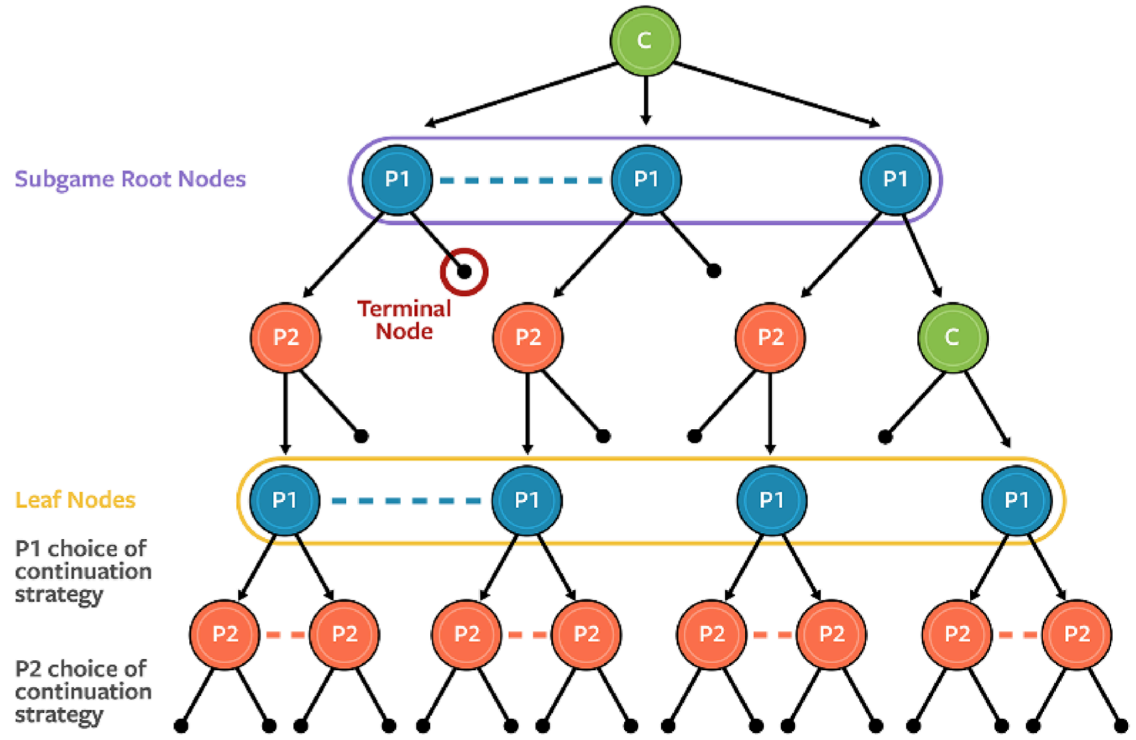
Imperfect Information:
Players can only observe *partial information* about the true underlying game state

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

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

Imperfect-Information Extensive-Form Games (IIEFGs)

[Kuhn 1953]



A commonly used formulation of games involving

- Multi-agent
- Sequential plays
- Imperfect information

💡 IIEFGs can be formulated as *Partially Observable Markov Games* (POMGs) with *tree structure* + *perfect recall* [Kovarik et al. 2019, Kozuno et al. 2021]

Image source: Superhuman AI for Multiplayer Poker, Brown & Sandholm 2019.

Two-Player Zero-Sum IIEFGs

Game value (expected cumulative reward):

$$V^{\mu,\nu} := \mathbb{E}\Big[\sum_{h=1}^{H} r_h(s_h, a_h, b_h) \mid a_h \sim \mu_h(\cdot \mid x_h), b_h \sim \nu_h(\cdot \mid y_h)\Big]$$

- μ: max-player
- ν: min-player
- $(x_h, y_h) = (x(s_h), y(s_h))$: *information sets* (observations) for the two players

Goal: Approximate Nash Equilibrium (controlling both players)

$$\mathrm{NEGap}(\mu,\nu) := \max_{\mu^\dagger} V^{\mu^\dagger,\nu} - \min_{\nu^\dagger} V^{\mu,\nu^\dagger} \le \varepsilon$$

Goal': No-regret (only control max player)

$$\mathrm{Reg}(T) := \max_{\mu^\dagger} \sum_{t=1}^{T} V^{\mu^\dagger,\nu^t} - V^{\mu^t,\nu^t} = o(T)$$

Online-to-batch conversion (e.g. [Zinkevich et al. 2007])

Play 2 no-regret algs against each other => Average policies are approximate Nash

Existing approaches

Full feedback / known game:

- Formulation as a linear program [von Stengel 1996, Koller et al. 1996, …]
- First-order optimization / online mirror descent (OMD) over sequence-form strategy space [Gilpin et al. 2008, Hoda et al. 2010, Kroer et al. 2015, Lee et al. 2021, …]
- Counterfactual regret minimization (CFR) [Zinkevich et al. 2007, Lanctot et al. 2009, Tammelin 2014, Burch et al. 2019, Farina et al. 2020b, …]

Bandit feedback (only observe trajectories from playing):

- Model-based approaches [Zhou et al. 2019, Zhang & Sandholm 2021]
- Monte-Carlo CFR (MCCFR) [Farina et al. 2020c, Farina & Sandholm 2021, …]
- Implicit-Exploration Online Mirror Descent (IXOMD) [Kozuno et al. 2021]
    - Learns an ε-Nash within $\widetilde{O}((X^2A + Y^2B)/\varepsilon^2)$ episodes (current best)
    - X, Y: number of information sets; A, B: number of actions
    - Lower bound is $\Omega((XA + YB)/\varepsilon^2)$, still $\max\{X, Y\}$ factor away

Question: How to design algorithms for learning Nash in two-player zero-sum IIEFGs from *bandit feedback* with *near-optimal sample complexity*?

Main Result

Theorem:
We design two new algorithms, **Balanced OMD** and **Balanced CFR**; both algorithms can learn an ε-Nash within $\widetilde{O}((XA + YB)/\varepsilon^2)$ episodes of play.

Algorithm	**OMD**	**CFR**	**Sample Complexity**
Zhang and Sandholm (2021)	- (model-based)		$\widetilde{\mathcal{O}}\left(S^2AB/\varepsilon^2\right)$
Farina and Sandholm (2021)		✓	$\widetilde{\mathcal{O}}(\mathrm{poly}\left(X, Y, A, B\right)/\varepsilon^4)$
Farina et al. (2021)	✓		$\widetilde{\mathcal{O}}\left(\left(X^4A^3 + Y^4B^3\right)/\varepsilon^2\right)$
Kozuno et al. (2021)	✓		$\widetilde{\mathcal{O}}\left(\left(X^2A + Y^2B\right)/\varepsilon^2\right)$
Balanced OMD (Algorithm 1)	✓		$\widetilde{\mathcal{O}}\left(\left(XA + YB\right)/\varepsilon^2\right)$
Balanced CFR (Algorithm 2)		✓	$\widetilde{\mathcal{O}}\left(\left(XA + YB\right)/\varepsilon^2\right)$
Lower bound (Theorem 6)	-	-	$\Omega\left(\left(XA + YB\right)/\varepsilon^2\right)$

Balanced OMD

Algorithm (Balanced OMD, max-player):

1. Play an episode with policy μ^t, construct loss estimator
$$\widetilde{\ell}^{\,t}_h(x_h, a_h) := \frac{\mathbf{1}\{(x_h^t, a_h^t) = (x_h, a_h)\} \cdot (1 - r_h^t)}{\mu_{1:h}^t(x_h, a_h) + \gamma \mu_{1:h}^{\star,h}(x_h, a_h)}.$$
2. Update policy
$$\mu^{t+1} = \operatorname*{argmin}_{\mu \in \Pi_{\max}} \eta \langle \widetilde{\ell}^{\,t}, \mu \rangle + D^{\mathrm{bal}}(\mu \| \mu^t),$$
(with efficient implementation)

Main new ingredient: **Balanced dilated KL distance**
$$D^{\mathrm{bal}}(\mu \| \nu) := \sum_{h, x_h, a_h} \frac{\mu_{1:h}(x_h, a_h)}{\mu_{1:h}^{\star,h}(x_h, a_h)} \log \frac{\mu_h(a_h \mid x_h)}{\nu_h(a_h \mid x_h)},$$

= Dilated KL [Hoda et al. 2010] + reweighting by **Balanced exploration policies**
$$\mu_{1:h}^{\star,h}(x_h, a_h) = \prod_{h'=1}^{h} \frac{|C_h(x_{h'}, a_{h'})|}{|C_h(x_{h'})|}$$

Number of descendants of $(x_{h'}, a_{h'})$ within h-th layer

(extension of [Farina et al. 2020c]).

Balanced OMD

Algorithm (Balanced OMD, max-player):

1. Play an episode with policy μ^t, construct loss estimator

$$\widetilde{\ell}^{\,t}_h(x_h, a_h) := \frac{\mathbf{1}\{(x_h^t, a_h^t) = (x_h, a_h)\} \cdot (1 - r_h^t)}{\mu_{1:h}^t(x_h, a_h) + \gamma \mu_{1:h}^{\star,h}(x_h, a_h)}.$$

2. Update policy

$$\mu^{t+1} = \operatorname*{argmin}_{\mu \in \Pi_{\max}} \eta \langle \widetilde{\ell}^{\,t}, \mu \rangle + D^{\mathrm{bal}}(\mu \| \mu^t),$$

(with efficient implementation)

Balanced OMD

Algorithm (Balanced OMD, max-player):

1. Play an episode with policy μ^t, construct loss estimator

$$\widetilde{\ell}^{\,t}_h(x_h, a_h) := \frac{\mathbf{1}\{(x_h^t, a_h^t) = (x_h, a_h)\} \cdot (1 - r_h^t)}{\mu^t_{1:h}(x_h, a_h) + \gamma \mu^{\star,h}_{1:h}(x_h, a_h)}.$$

2. Update policy

$$\mu^{t+1} = \operatorname*{argmin}_{\mu \in \Pi_{\max}} \eta \langle \widetilde{\ell}^{\,t}, \mu \rangle + D^{\mathrm{bal}}(\mu \| \mu^t),$$

(with efficient implementation)

Theorem: Balanced OMD achieves regret bound

$$\mathrm{Reg}(T) \le \widetilde{O}(\sqrt{H^3 XAT})$$

and learns ε-Nash within $\widetilde{O}(H^3(XA + YB)/\varepsilon^2)$ episodes of self-play.

Balanced OMD

Algorithm (Balanced OMD, max-player):

1. Play an episode with policy μ^t, construct loss estimator

$$\widetilde{\ell}^{\,t}_h(x_h, a_h) := \frac{\mathbf{1}\{(x_h^t, a_h^t) = (x_h, a_h)\} \cdot (1 - r_h^t)}{\mu^t_{1:h}(x_h, a_h) + \gamma \mu^{\star,h}_{1:h}(x_h, a_h)}.$$

2. Update policy

$$\mu^{t+1} = \operatorname*{argmin}_{\mu \in \Pi_{\max}} \eta \langle \widetilde{\ell}^{\,t}, \mu \rangle + D^{\mathrm{bal}}(\mu \| \mu^t),$$

   (with efficient implementation)

Theorem: Balanced OMD achieves regret bound

$$\mathrm{Reg}(T) \le \widetilde{O}(\sqrt{H^3 XAT})$$

and learns ε-Nash within $\widetilde{O}(H^3(XA + YB)/\varepsilon^2)$ episodes of self-play.

Main technical highlight:

"Balancing effect" introduced by D^{bal} (adapts to geometry of policy space)

==> better stability bound than existing OMD analyses (e.g. [Kozuno et al. 2021]) ,

by bounding a certain *log-partition function* via *2nd order Taylor expansion*

Balanced CFR

Mixture of $\mu^{\star,h}$ and μ^t

Algorithm (Balanced CFR, max-player):

1. Play **H** episodes with policy $\mu_{1:h}^{\star,h}\mu_{h+1:H}^t$, observe trajectory
$$(x_1^{t,(h)}, a_1^{t,(h)}, r_1^{t,(h)}, \ldots, x_H^{t,(h)}, a_H^{t,(h)}, r_H^{t,(h)})$$
2. Construct counterfactual loss estimator
$$\widetilde{L}_h^t(x_h, a_h) := \frac{\mathbf{1}\{(x_h^{t,(h)}, a_h^{t,(h)}) = (x_h, a_h)\}}{\mu_{1:h}^{\star,h}(x_h, a_h)} \cdot \sum_{h'=h}^{H} (1 - r_{h'}^{t,(h)}).$$
3. Update policy at each information set via **Hedge**
$$\mu_h^{t+1}(a \mid x_h) \propto_a \mu_h^t(a \mid x_h) \cdot \exp\Big(-\eta\mu_{1:h}^{\star,h}(x_h, a_h)\widetilde{L}_h^t(x_h, a_h)\Big).$$
(can also use Regret Matching [Zinkevich et al. 2007].)

Algorithm =

MCCFR framework [Lanctot et al. 2009, Farina et al. 2020c]

+ sampling by **mixing importance weighting** (using $\mu^{\star,h}$) **and Monte Carlo** (using μ^t)

+ "adaptive" learning rate $\mu_{1:h}^{\star,h}(x_h, a_h)$ at each infoset

Balanced CFR

Mixture of $\mu^{\star,h}$ and μ^{t}

Algorithm (Balanced CFR, max-player):

1. Play **H** episodes with policy $\mu^{\star,h}_{1:h}\mu^{t}_{h+1:H}$, observe trajectory

$$(x_1^{t,(h)}, a_1^{t,(h)}, r_1^{t,(h)}, \ldots, x_H^{t,(h)}, a_H^{t,(h)}, r_H^{t,(h)})$$

2. Construct counterfactual loss estimator

$$\widetilde{L}_h^t(x_h, a_h) := \frac{\mathbf{1}\{(x_h^{t,(h)}, a_h^{t,(h)}) = (x_h, a_h)\}}{\mu_{1:h}^{\star,h}(x_h, a_h)} \cdot \sum_{h'=h}^{H} (1 - r_{h'}^{t,(h)}).$$

3. Update policy at each information set via **Hedge**

$$\mu_h^{t+1}(a \,|\, x_h) \propto_a \mu_h^t(a \,|\, x_h) \cdot \exp\Big(-\eta \mu_{1:h}^{\star,h}(x_h, a_h) \widetilde{L}_h^t(x_h, a_h) \Big).$$

(can also use Regret Matching [Zinkevich et al. 2007].)

Balanced CFR

Algorithm (Balanced CFR, max-player):

Mixture of $\mu^{\star,h}$ and μ^t

1. Play **H** episodes with policy $\mu^{\star,h}_{1:h}\mu^t_{h+1:H}$, observe trajectory

$$(x_1^{t,(h)}, a_1^{t,(h)}, r_1^{t,(h)}, \ldots, x_H^{t,(h)}, a_H^{t,(h)}, r_H^{t,(h)})$$

2. Construct counterfactual loss estimator

$$\widetilde{L}_h^t(x_h, a_h) := \frac{\mathbf{1}\{(x_h^{t,(h)}, a_h^{t,(h)}) = (x_h, a_h)\}}{\mu_{1:h}^{\star,h}(x_h, a_h)} \cdot \sum_{h'=h}^{H} (1 - r_{h'}^{t,(h)}).$$

3. Update policy at each information set via **Hedge**

$$\mu_h^{t+1}(a \mid x_h) \propto_a \mu_h^t(a \mid x_h) \cdot \exp\Big(- \eta \mu_{1:h}^{\star,h}(x_h, a_h) \widetilde{L}_h^t(x_h, a_h) \Big).$$

(can also use Regret Matching [Zinkevich et al. 2007].)

Theorem: Balanced CFR learns ε-Nash within $\widetilde{O}(H^4(XA + YB)/\varepsilon^2)$ episodes of self-play.

🤔 $\{\mu^t\}_{t=1}^T$ also achieves $\mathrm{Reg}(T) \leq \widetilde{O}(\sqrt{H^3XAT})$, but are not actual played policies.

Main technical highlight:

Sharp counterfactual regret decomposition + reduced variance brought by $\mu^{\star,h}$

Coarse Correlated Equilibria (CCEs) in multi-player IIEFGs

Normal-Form Coarse Correlated Equilibrium

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

No gains in deviating from *correlated policy* $\boldsymbol{\pi}$

Corollary: Run Balanced OMD or Balanced CFR on all players ==> ε-NFCCE of multi-player general-sum IIEFGs within $\widetilde{O}((\max_i X_i A_i)/\varepsilon^2)$ episodes of play.

Proof follows directly by known connection between NFCCE and no-regret learning in multi-player general-sum IIEFGs [Celli et al. 2019].

Summary

First line of near-optimal algorithms for learning IIEFGs from bandit feedback

Crucial use of **balanced exploration policies**

- distance functions in OMD
- sampling policies in CFR

Future directions

- Further understandings of OMD/CFR type algorithms
- Sample-efficient learning of other equilibria (e.g. correlated equilibria)
- Relationship between Markov Games and Extensive-Form Games
- Empirical investigations

Thank you!

Paper: https://arxiv.org/abs/2202.01752