

Actor-Critic Policy Optimization in a Large-Scale Imperfect-Information Game

Haobo Fu et al.

Tencent AI Lab, Shenzhen, China

haobofu@tencent.com

ICLR-2022

Two-player Zero-sum Imperfect-Information Games (IIGs)

An optimal solution to a 2-player zero-sum IIG usually refers to a **Nash Equilibrium** (NE), where no player could improve by unilaterally deviating to a different policy.

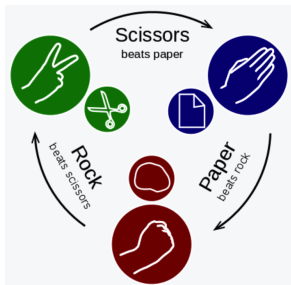


Figure: For instance, in the 2-player Rock-Paper-Scissors game, the NE is for both players playing the Uniform random policy: $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$.

Single-agent Reinforcement Learning (RL) Methods

- Single-agent RL methods with self-play for 2-player zero-sum IIGs **do not converge to a NE**, because learning becomes **non-stationary** and **non-Markovian** when multiple agents learn simultaneously in a competitive environment.

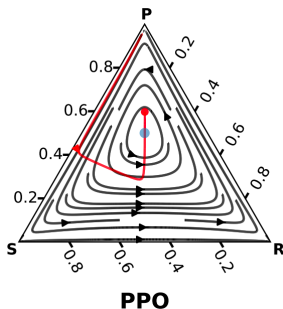
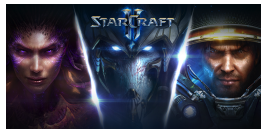


Figure: For instance, in a biased 2-player Rock-Paper-Scissors game, PPO [12] with self-play does not converge to the NE, which is denoted by the **blue** point.

Single-agent RL Methods

- Nonetheless, single-agent RL methods with self-play are **efficiently scalable**, with successful applications to very large-scale games:



(a) Starcraft II [16]



(b) Dota 2 [1]




(c) Honor of Kings [18]

Figure: Successful applications of single-agent RL methods to large-scale IIGs.

Counterfactual Regret Minimization (CFR)

- CFR [19] is a **tabular** iterative algorithm that minimizes the total regret of strategies by minimizing the cumulative counterfactual regret in every state (infoset).
- The **average policy** in CFR is **guaranteed to converge to a NE** in two-player zero-sum IIGs.
- To handle large-scale IIGs with CFR, **abstractions** (applied to either the action space or the state space) are usually employed to reduce the game to a manageable size [11, 3, 4]¹.
- However, abstractions are usually **domain specific** [17, 8, 6]. More importantly, some large-scale IIGs are **inherently difficult to be abstracted**, such as the game of 1v1 Mahjong investigated in this paper.

¹DeepStack [11] employs sparse lookahead trees, much like the action abstraction. 

Our New Neural Extension to CFR: NW-CFR

Neural-based Weighted CFR (NW-CFR)

- employs a neural network to generalize across states and relies on **only trajectory samples** for training.
- approximates the expectation of the sum of **sampled advantages** $R_t^a(s, a) := \mathbb{E}[\sum_{k=1}^t \tilde{A}^{\pi_k}(s, a)]$.
- is a straightforward neural extension to a type of **weighted CFR**, which is defined as:

Definition (Weighted CFR)

Weighted CFR follows the same procedure as the original CFR [19], except that the instantaneous counterfactual regret $r_t^c(s, a)$ is weighted by some weight $w_t(s)$, $w_t(s) > 0$ and $\sum_{t=0}^{\infty} w_t(s) = \infty$. The original CFR is a type of weighted CFR with $w_t(s) = 1.0$.

A Major Difference Between NW-CFR & Others [2, 9, 15]

- NW-CFR operates on the **sampled advantage** $\tilde{A}^{\pi_k}(s, a)$ instead of the **sampled instantaneous counterfactual regret** $\tilde{r}_k^c(s, a)$, and we have

$$\tilde{r}_k^c(s, a) = [f_p^{\mu_k}(s)]^{-1} \tilde{A}^{\pi_k}(s, a), \quad (1)$$

where $f_p^{\mu_k}(s)$ is a probability of reaching state s .

- The **larger variance** (due to $[f_p^{\mu_k}(s)]^{-1}$) of $\tilde{r}_k^c(s, a)$, compared with $\tilde{A}^{\pi_k}(s, a)$, may have a negative influence on the performance when function approximation is used with only trajectory samples. This influence is magnified in games with long episodes and large info set size.

Theoretical Properties of NW-CFR and Weighted CFR

Theorem (NW-CFR and weighted CFR)

NW-CFR is equivalent to a type of weighted CFR with Hedge when $w_t(s) = f_p^{\mu_t}(s) > 0$, given that enough trajectories are sampled and $y(a|s; \theta_t)$ is sufficiently close to $R_t^a(s, a)$. Further, if $\eta(s) = \sqrt{8 \ln |\mathcal{A}(s)| / \{[w_h(s)]^2 \Delta^2(s) T\}}$ and $w_t(s) = f_p^{\mu_t}(s) \in [w_l(s), w_h(s)] \subset (0, 1]$, $t = 1, \dots, T$, the average policy^a $\bar{\pi}$ of the corresponding weighted CFR with Hedge and equivalently NW-CFR with $\bar{\pi}_p(a|s) = \sum_{t=1}^T [f_p^{\pi_t}(s) \pi_t(a|s)] / \sum_{t=1}^T f_p^{\pi_t}(s)$, $\forall p \in \mathcal{P}$, has ϵ exploitability after T iterations, where

$$\epsilon \leq |S| \Delta \sqrt{\frac{1}{2T} \ln |\mathcal{A}|} + \Delta \sum_{s \in S} \frac{w_h(s) - w_l(s)}{w_h(s)}. \quad (2)$$

^aGiven $\pi_{p,t}$ at each iteration, we could obtain $\bar{\pi}_p$ using the techniques introduced in [14].

ACH²: a Practical Implementation of NW-CFR

Our new algorithm Actor-Critic Hedge (ACH)

- employs a framework of decoupled acting and learning (similar to IMPALA [5]) and thus is **efficiently scalable**.
- trains **the current policy** with an entropy regularization on only sampled states, without the calculation of the average policy.
- uses only trajectory samples at the current iteration and requires no computation of best response, thus having **similar complexity to RL-style methods**, e.g., PPO.

²The code is available at https://github.com/Liuweiming/ACH_poker

Our Proposed 1-on-1 Mahjong Benchmark³

- is **the first 1-on-1 Mahjong benchmark**, and the corresponding game is played widely in Tencent online games. (<https://maji.ang.qq.com>)
- has **a longer game length** than poker, which makes sampling multiple actions in a state prohibitive.
- has **a larger info set size** than poker (as shown below), which may be more difficult for methods relying on only trajectory samples.

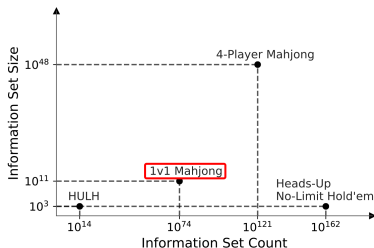
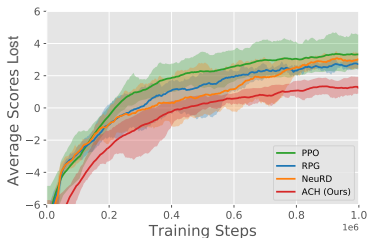


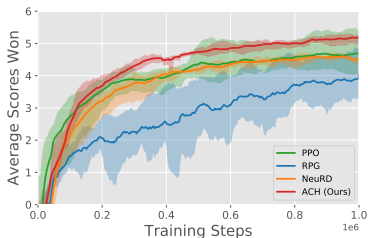
Figure: The game complexity of Mahjong and poker.

³The code is available at <https://github.com/yata0/Mahjong>

Experimental Results on 1-on-1 Mahjong



(a) Approximate Lower Bound Exploitability



(b) Head-to-Head performance

Figure: (a): The training curves of the best response against each agent. **Lower is better.** (b): The training curves of each agent. The performance of an agent is evaluated by the average scores the agent wins against a common rule-based agent. **Higher is better.** ACH is **significantly more difficult to be exploited** than other RL-style methods (PPO, RPG [13], and NeuRD [7]).

Experimental Results on 1-on-1 Mahjong

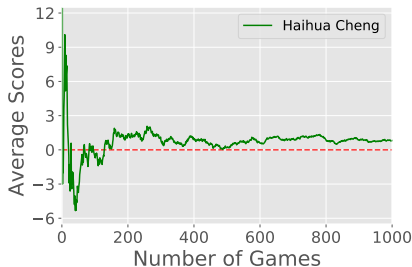


Figure: After playing 1,000 games, **JueJong** (the agent obtained by ACH) **won the champion** by a score of 0.82 ± 0.96 (mean \pm standard deviation), with a p-value of 0.19 under one-tailed t-test. Hence, we may conclude that Haihua Cheng failed to exploit JueJong effectively within 1,000 games.

Experimental Results on 1-on-1 Flop Hold'em Poker (FHP)

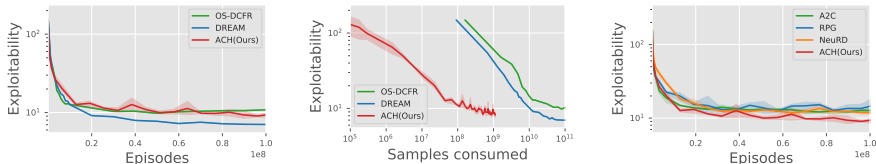


Figure: The exploitability on FHP, with the x-axis being the number of episodes generated (**left** and **right**) and the number of samples consumed (**middle**). ACH achieves an exploitability of 10 chips/game (a big blind is 100 chips) almost **100** times faster than DREAM [15] and **1,000** times faster than OS-DCFR [2]. Also, ACH converges **significantly faster** and **achieves a lower exploitability**, in comparison with other RL-style algorithms (A2C [10], RPG, and NeuRD).

- A new model-free actor-critic algorithm, i.e., **ACH**, for approximating a NE in large-scale IIGs is developed.
- ACH is a practical implementation of a new neural-based CFR algorithm, i.e., **NW-CFR**, which has a theoretical guarantee and operates on the **sampled advantage**, which has a **smaller variance** than the sampled instantaneous counterfactual regret.
- The first **1-on-1 Mahjong benchmark** is introduced.
- JueJong, i.e., the agent obtained by ACH, **defeats a human champion** on 1-on-1 Mahjong.

Thank you for your attention!

References I

- [1] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Christopher Hesse, and et al.
Dota 2 with large scale deep reinforcement learning.
CoRR, abs/1912.06680, 2019.
- [2] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm.
Deep counterfactual regret minimization.
In *International Conference on Machine Learning (ICML)*, pages 793–802, 2019.
- [3] Noam Brown and Tuomas Sandholm.
Superhuman AI for heads-up no-limit poker: Libratus beats top professionals.
Science, 359(6374):418–424, 2018.

References II

- [4] Noam Brown and Tuomas Sandholm.
Superhuman AI for multiplayer poker.
Science, 365(6456):885–890, 2019.
- [5] Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan,
Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley,
Iain Dunning, Shane Legg, and Koray Kavukcuoglu.
IMPALA: scalable distributed deep-rl with importance weighted
actor-learner architectures.
In International Conference on Machine Learning (ICML), volume 80,
pages 1406–1415, 2018.
- [6] Sam Ganzfried and Tuomas Sandholm.
Potential-aware imperfect-recall abstraction with earth mover’s
distance in imperfect-information games.
In Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

- [7] Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Rémi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duéñez-Guzmán, et al.
Neural replicator dynamics: Multiagent learning via hedging policy gradients.
In International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS), pages 492–501, 2020.
- [8] Michael Johanson, Neil Burch, Richard Valenzano, and Michael Bowling.
Evaluating state-space abstractions in extensive-form games.
In Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems, pages 271–278, 2013.

- [9] Hui Li, Kailiang Hu, Shaohua Zhang, Yuan Qi, and Le Song.
Double neural counterfactual regret minimization.
In *International Conference on Learning Representations (ICLR)*,
2020.
- [10] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex
Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray
Kavukcuoglu.
Asynchronous methods for deep reinforcement learning.
In *International Conference on Machine Learning (ICML)*, pages
1928–1937, 2016.

- [11] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling.
Deepstack: Expert-level artificial intelligence in heads-up no-limit poker.
Science, 356(6337):508–513, 2017.
- [12] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov.
Proximal policy optimization algorithms.
CoRR, abs/1707.06347, 2017.

- [13] Sriram Srinivasan, Marc Lanctot, Vinicius Zambaldi, Julien Pérolat, Karl Tuyls, Rémi Munos, and Michael Bowling.
Actor-critic policy optimization in partially observable multiagent environments.
In Advances in Neural Information Processing Systems (NeurIPS), pages 3422–3435, 2018.
- [14] Eric Steinberger.
Single deep counterfactual regret minimization.
arXiv preprint arXiv:1901.07621, 2019.
- [15] Eric Steinberger, Adam Lerer, and Noam Brown.
DREAM: deep regret minimization with advantage baselines and model-free learning.
CoRR, abs/2006.10410, 2020.

- [16] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, and et al. John P. Agapiou and.
Grandmaster level in StarCraft II using multi-agent reinforcement learning.
Nature, 575(7782):350–354, 2019.
- [17] Kevin Waugh, Martin Zinkevich, Michael Johanson, Morgan Kan, David Schnizlein, and Michael Bowling.
A practical use of imperfect recall.
In *Eighth symposium on abstraction, reformulation, and approximation*, 2009.

- [18] Deheng Ye, Guibin Chen, Wen Zhang, Sheng Chen, Bo Yuan, Bo Liu, Jia Chen, Zhao Liu, Fuhao Qiu, Hongsheng Yu, Yinyuting Yin, Bei Shi, Liang Wang, Tengfei Shi, Qiang Fu, Wei Yang, Lanxiao Huang, and Wei Liu.
Towards playing full MOBA games with deep reinforcement learning.
In Advances in Neural Information Processing Systems (NeurIPS), 2020.
- [19] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione.
Regret minimization in games with incomplete information.
In Advances in Neural Information Processing Systems (NeurIPS), pages 1729–1736, 2008.