PT-DCFR: Accelerating and Improving Deep CFR Using Population Based Training (Student Abstract)

Dingzhong Cai[1], Huale Li[2*], Hang Xiao[1], Shuhan Qi[3†], Jiajia Zhang[3]
[1]Northwestern Polytechnical University
[2]Lanzhou University
[3]Harbin Institute of Technology, Shenzhen
caidz@mail.nwpu.edu.cn, lihuale@lzu.edu.cn, shawh@mail.nwpu.edu.cn, shuhanqi@cs.hitsz.edu.cn, zhangjiajia@hit.edu.cn

Abstract

Deep CFR enables end-to-end approximation of Nash equilibria in imperfect-information games(IIGs) but is sensitive to hyperparameters, making manual tuning inefficient. To address this, we propose PT-DCFR, which integrates Population-Based Training (PBT) with Deep CFR to dynamically optimize hyperparameters during training. Building upon this, we further introduce P2T-DCFR, which decouples parameter selection from model performance.

Introduction

Artificial intelligence has demonstrated remarkable success in perfect-information games such as Go (Silver et al. 2016). However, IIGs, which more closely resemble real-world decision-making scenarios, present continuing challenges. Counterfactual Regret Minimization (CFR) (Zinkevich et al. 2007) and its variants(Li et al. 2024b) form a foundational methodology for solving IIGs. A primary constraint of these approaches is their dependence on abstraction to handle large games, which introduces expert dependency and information loss.

Deep CFR (Brown et al. 2019) overcame this limitation by employing deep neural networks to approximate strategies end-to-end. Subsequent methods including SD-CFR (Steinberger 2019), D2CFR (Li et al. 2024a) and OD-CFR(Wang et al. 2025) extended this line of work. However, these deep learning algorithms are highly sensitive to hyperparameter settings, and their manual tuning is often inefficient and computationally expensive.

To address these challenges, we propose PT-DCFR and its extension P2T-DCFR. PT-DCFR integrates PBT(Jaderberg et al. 2017) with Deep CFR, using parallel agent training and experience sharing to enhance data diversity and enable dynamic hyperparameter optimization. P2T-DCFR further introduces a hyperparameter potential metric to decouple parameter assessment from agent performance, preventing premature discarding of promising configurations. Our methods demonstrate significantly lower exploitability and faster convergence than Deep CFR, with additional experiments confirming generalizability to other algorithms like D2CFR.

*Corresponding author
†Corresponding author

Methodology

PT-DCFR integrates PBT to dynamically optimize hyperparameters during strategy learning. A population of n agents is maintained and trained in parallel. After every δ iterations, each agent is evaluated based on its strategy's exploitability. Agents are ranked, and the bottom τ agents inherit the network weights and hyperparameters from the top τ, with hyperparameters perturbed to preserve diversity while exploring better combinations.

To accommodate the periodic evaluation required by PBT, we modify the policy network training scheme. Rather than training only at the end as in Deep CFR, the policy network is trained incrementally after each δ iterations using samples from the global samples pool $\tilde{B}^v$. To manage computational cost, the number of training steps per session is reduced to $\frac{\delta}{T}$ of the original.

A key advantage of the multi-agent framework is the enhanced diversity and stability of training samples. Each agent independently generates distinct trajectories, all of which are aggregated into a global pool $\tilde{B}^v$. This shared experience pool enriches the training data for every agent, leading to more robust and efficient learning.

In terms of computational resource, the total number of traversals and value network updates remains asymptotically equivalent to that of Deep CFR. The additional overhead primarily arises from: (1) periodic policy network training and (2) strategy evaluation during population ranking. Although PT-DCFR requires moderately more memory and computation, the parallelized training structure significantly reduces wall-clock time. Coupled with the substantial improvement in strategy quality, this overhead is well justified.

PT-DCFR couples hyperparameter selection to immediate agent performance, which can overlook hyperparameters with strong long-term potential. P2T-DCFR addresses this by introducing a novel evaluation metric: *hyperparameter potential*.

The hyperparameter potential p for agent x_i at generation g is defined as:

$$x_i^g(p) = \beta(x_i^g(s) - x_i^{g-1}(s)) + x_{i-1}^g(p) \quad (1)$$

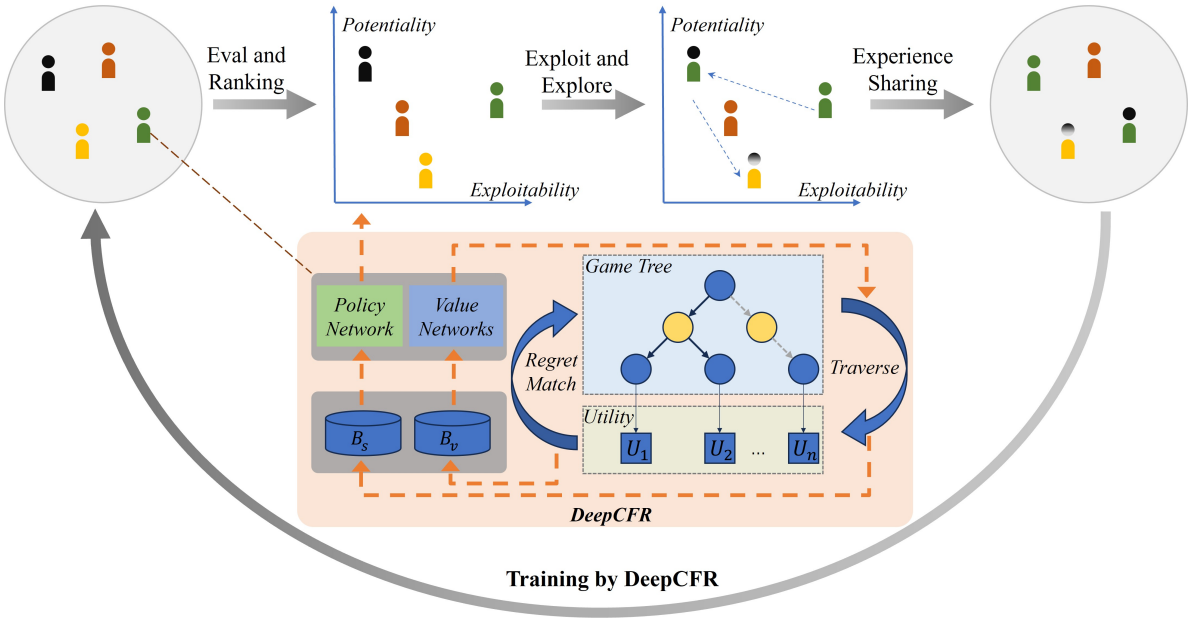

Figure 1: The overall framework of P2T-DCFR. Hyperparameters and network weights are denoted by ● and ▮ respectively.

	Deep CFR Variants				D2CFR Variants		
Game	Deep CFR	SD-CFR	PT-DCFR	P2T-DCFR	D2CFR	PT-D2CFR	P2T-D2CFR
Leduc	17.8±4.5	11.2±1.5	14.3±1.9	**9.8±2.1**	14.5±2.2	10.1±1.2	**8.4±0.9**
Leduc(8)	18.5±2.2	13.4±0.7	15.0±0.7	**12.2±0.4**	14.5±1.5	14.6±1.3	**12.6±0.9**
Leduc(10)	15.1±2.1	14.1±0.5	12.2±1.0	**10.3±0.7**	12.1±1.0	11.0±0.9	**10.6±0.7**
Liar's Dice	9.5±0.6	7.3±0.4	8.7±0.5	**6.3±1.2**	9.7±0.6	8.0±0.6	**6.7±0.5**

Table 1: Comparison of exploitability (10^{-2}) across different Deep CFR and D2CFR variants.

$$x_i^g(s) = \frac{1}{x_i^g(e)^2} \tag{2}$$

where $x_i^g(s)$ is the agent's score (derived from its exploitability e), and β is a discount factor.

P2T-DCFR incorporates a two-stage evaluation process: model performance and hyperparameter potential. Agents are first ranked according to model score for model exploitation. Agents are then ranked by hyperparameter potential to guide hyperparameter exploration. After hyperparameter exploitation and exploration, the historical potential value for the affected agent is reset to 0. This decouples the quality of the hyperparameter from the transient performance of the coupled model, enabling a more objective selection of hyperparameters with high long-term value.

Experiments

We evaluated the performance of Deep CFR, SD-CFR, PT-DCFR and P2T-DCFR in four games: standard 6-card Leduc Poker, its extended variants Leduc(8) and Leduc(10) with 8 and 10 cards respectively, and Liar's Dice.

As shown in Table 1, P2T-DCFR achieves the lowest exploitability among the Deep CFR variants in all tested games, demonstrating the superior convergence and optimization efficacy of our proposed hyperparameter optimization framework. In particular, the P2T-based framework also generalizes effectively to D2CFR, where P2T-D2CFR consistently outperforms both D2CFR and PT-D2CFR. These results collectively confirm that our PBT approach with dynamic hyperparameter optimization significantly enhances strategy precision, while maintaining strong adaptability across different base algorithms.

To validate the effectiveness of experience sharing, we compare the Deep CFR against a variant employing parallel sampling and experience sharing (PSES-Deep CFR). Under an identical computational budget, PSES-Deep CFR reduces exploitability on Leduc from **0.100** to **0.060**, underscoring that diversifying training data through parallel sampling is highly effective.

Conclusion

We introduced PT-DCFR and P2T-DCFR, which enhance Deep CFR through population-based hyperparameter optimization and a novel potential metric. Our methods improve training efficiency and strategy accuracy via dynamic hyperparameter tuning and experience sharing. Experimental results demonstrate significant improvements in convergence and performance across multiple imperfect-information games.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (No.62406251).

References

Brown, N.; Lerer, A.; Gross, S.; and Sandholm, T. 2019. Deep counterfactual regret minimization. In *International conference on machine learning*, 793–802. PMLR.

Jaderberg, M.; Dalibard, V.; Osindero, S.; Czarnecki, W. M.; Donahue, J.; Razavi, A.; Vinyals, O.; Green, T.; Dunning, I.; Simonyan, K.; et al. 2017. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*.

Li, H.; Wang, X.; Guo, Z.; Zhang, J.; and Qi, S. 2024a. D2CFR: Minimize Counterfactual Regret With Deep Dueling Neural Network. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12): 18343–18356.

Li, K.; Xu, H.; Fu, H.; Fu, Q.; and Xing, J. 2024b. Automatically designing counterfactual regret minimization algorithms for solving imperfect-information games. *Artificial Intelligence*, 337: 104232.

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(7587): 484–489.

Steinberger, E. 2019. Single deep counterfactual regret minimization. *arXiv preprint arXiv:1901.07621*.

Wang, J.; Li, Y.; Niu, S.; and Wu, Z. 2025. On deep CFR integrated with opponent model in imperfect information games. *Knowledge-Based Systems*, 327: 114105.

Zinkevich, M.; Johanson, M.; Bowling, M.; and Piccione, C. 2007. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20.