Multi-Agent Counterfactual Regret Minimization for Partial-Information Collaborative Games

Matthew Hartley
Caltech
mhartley@caltech.edu

Stephan Zheng
Caltech
stephan@caltech.edu

Yisong Yue
Caltech
yyue@caltech.edu

Abstract

We study the generalization of counterfactual regret minimization (CFR) to partial-information collaborative games with more than 2 players. For instance, many 4-player card games are structured as 2v2 games, with each player only knowing the contents of their own hand. To study this setting, we propose a multi-agent collaborative version of Kuhn Poker. We observe that a straightforward application of CFR to this setting can lead to sub-optimal results, and explore extensions to CFR that can offer improved performance.

Counterfactual Regret Minimization (CFR) is an iterative learning approach for multi-agent adversarial partial-information games. The goal of CFR is to iteratively minimize a regret bound, called counterfactual regret, on the utility of different actions. Since counterfactual regret is an upper bound on the true regret, CFR also minimizes true regret. For two player zero-sum games (e.g., head's up poker), CFR therefore converges to the unique Nash equilibrium (Zinkevich et al. [2008]). Recently, Moravčík et al. [2017] combined CFR with state-space compression via deep learning and showed this to be effective in beating human players at 2-player no limit Texas Hold'em poker.

In this paper, we study the generalization of CFR to partial-information games with *more than 2 players*. In such games, the player dynamics can be much richer, e.g. an optimal strategy might require players to both collaborate and compete. For instance, many 4-player card games are structured as 2v2 games, with each player only knowing the contents of their own hand. Canonical examples include Spades, Euchre, and Bridge. In general, it is not known whether Nash equilibria strategies exist in these games, or whether CFR can converge to good solutions.

To study this question, we developed a collaborative extension of Kuhn Poker, which can be viewed as arguably the simplest game in this setting that admits interesting strategic behavior. We compare tabular CFR with extensions where 1) agents attempt to maximize their team's expected utility and 2) players are given information about their partner's hands through a noisy channel, to simulate how adding a state inference model could help agents play. Related reinforcement learning approaches were studied in Foerster et al. [2017], which found collaborative strategies can be learned by explicitly including the response of other players into policy updates.

Our contributions are as follows. We show that when applying CFR in the 4-player setting:

- Basic CFR learns bad strategies, where *allies are antagonistic, rather than collaborative*.
- Using collective rewards yields strategies that dominate those found with selfish rewards.
- We analyze the performance of CFR with an additional state inference oracle that reveals hidden state information. We find that players with a perfect oracle learn a strategy that dominates baseline players without oracle.
- However, we find that CFR is *not robust*: close-to-perfect oracles significantly degrade the quality of learned strategies, making them worse than strategies that do not use an oracle. This result implies there is a nontrivial performance bound on which oracles are useful.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

1 Collaborative Kuhn Poker

To study extensions of CFR, we propose *Collaborative Kuhn Poker (CKP)*, a 4-player extension of Kuhn Poker (Kuhn [1950]). CKP uses a deck of 6 cards: Queen, King and Ace, of Hearts or Spades. There are 4 players $i \in$ [North, West, South and East] and 2 teams: North-South and East-West.

Each game round, the players are given $N = 3$ chips and a private card s^i; the private cards are sampled without replacement from the deck. At the start of a round, each player places one chip in the pot. A round then proceeds in turns; each turn t, players take an action a_t: betting a chip, raising the bet by 1, calling the bet, or folding. If any player folds, their partner automatically folds and the other team calls the outstanding bet. Once a bet is called or if neither team folds, the team with the strongest poker hand (neglecting flushes) wins the pot, which is shared equally within the team.

2 Generalizing Counterfactural Regret Minimization

CFR Partial-information games can be formalized by *infosets* I_t^i (all information known to player i at time t) and a strategy profile, which encodes player behavior as a map $\sigma : I \mapsto P(a|I)$ from infosets to distributions over the possible actions a. To learn the optimal strategy σ^*, CFR algorithms minimize the counterfactual regret $R_i(I, a; \sigma)$, which is defined as:

$$R_i(I, a) = \frac{1}{T} \sum_t \pi_{-i}^{\sigma^t}(I) \times (U(\sigma^t | I \to a, I) - U(\sigma^t, I)) \tag{1}$$

where $\pi_{-i}^{\sigma^t}$ is the probability of reaching infoset I given a strategy profile σ^t with the exception being that the current player's strategy was to reach the given infoset, $U(\sigma, I)$ is the expected reward of using strategy σ^t at infoset I, and $\sigma | I \to a$ is the strategy that is identical to σ, except for that the player always makes action a at infoset I. CFR now learns σ^* by iteratively updating:

$$\forall I, a: \quad \sigma^{k+1}(I, a) = \begin{cases} \frac{R_i^{k+}(I,a)}{\sum_a R_i^{k+}(I,a)}, & \sum_a R_i^{k+}(I, a) > 0 \\ \frac{1}{|A|}, & \sum_a R_i^{k+}(I, a) \leq 0 \end{cases} \tag{2}$$

where R^k is the accumulated regret up to iteration k, $R_i^{k+} = \max(R_i^k, 0)$ and the initial σ^0 is random. Zinkevich et al. [2008] showed that the average strategy $\frac{1}{K} \sum_{k=1}^K \sigma^k \to \sigma^*$ then converges to an ϵ-Nash equilibrium, which implies CFR finds a near-optimal strategy in 2-player zero-sum games.

Collective Rewards In 4-player collaborative games, one way to force the algorithm to converge to a collaborative solution is to design an appropriate reward function. The simplest version is to set the reward for each player be the combined reward of their entire team. Such a reward function discourages making an action that is in a player's self interest, but not the team's. Collaborative reward structures share affinity to real collaborative games such as Bridge, which has shared rewards for both players in the team.

State Inference Model A core issue in partial-information games is how to reduce uncertainty over the unseen state information, i.e. the cards in the other players' hands. Therefore, we study the effect of (noisy) state inference models ("oracle") on CFR in the 4-player setting. An oracle simulates increasingly informative guesses of other players' cards as a round unfolds, which can support decisions. For instance, in CKP a player could fold their partner's raise when they have a weak card (e.g. a Queen), suggesting they have a high likelihood to lose. However, an oracle could reveal that their joint hand is strong (e.g. their partner also has a Queen, giving the team a pair). In other contexts, attempting to model unknown data has helped in finding better policies in partial information setting (Barrett et al. [2013]) so it is plausible that such a model would help in this case.

For the sake of simplicity, our oracle outputs a player's partner's card with a given probability p, and outputs a uniform distribution of cards with probability $1 - p$. To simulate the extra information that the players' actions reveal during a game, we increase from $p = 0$ (completely uninformative) towards $p = 0.9$ (almost perfect) in 0.25 increments during each round.[1]

[1]Note that perfect state inference is almost always impossible in practice.

Table 1: Expected rewards of CFR and its extensions.

Algorithm	North	East	South	West	Inter-team Δ	Intra-team δ
`CFR`	-0.071	-0.200	$+0.063$	$+0.208$	$+0.008$	$+0.270$
`CFR-tied-`r	-0.02	-0.161	$+0.125$	$+0.056$	$+0.105$	$+0.181$
`CFR+oracle`	-0.111	$+0.008$	$+0.026$	$+0.077$	$+0.086$	$+0.102$
`CFR+oracle-`p	-0.05	-0.04	$+0.21$	$+0.00$	$+0.10$	$+0.15$

Table 2: Pairwise advantage, computed as gains for the column model.

Matchup	`CFR-tied-`r	`CFR+oracle`	`CFR+oracle-`p
`CFR`	$+0.788$	$+0.335$	-1.1
`CFR-tied-`r		$+0.498$	-1.8

3 Empirical Evaluation

We trained all models using (2), using a batch of 360 hands for each update. All models were trained for 1,000 batches (360,000 hands), after which we evaluated the final average strategy (see Section 2).

In order to quantify the level of collaboration, we compared how many coins transferred 1) between teams ("intra") to 2) that within teams ("inter"). The intra-team Δ and inter-team transfer δ is:

$$\Delta = \Delta_N + \Delta_S, \quad \Delta_i = \mathbb{E}[r^i] - r_0^i, \quad \delta_{ij} = \mathbb{E}[r^i] - \mathbb{E}[r^j], \tag{3}$$

where r_0^i is the starting stack of chips for player i and we use the zero-sum property $\Delta_N + \Delta_S = -(\Delta_E + \Delta_W)$: North-South's gain is East-West's loss. Strategies with high intra-team and low inter-team transfer can indicate defection, e.g. when a player folds when their partner has raised.

To evaluate the quality of the learned strategies, we calculated their expected reward and pairwise advantages in all possible model match-ups. To determine the pairwise advantage of two strategies σ_i and σ_j, we calculated the expected rewards of σ_i (σ_j) in North-South (East-West) and vice versa. The pairwise advantage is the difference between the two results.

CFR We found that CKP is not well solved by CFR. When using batch-learning (the policy is updated after seeing a minibatch of hands), the algorithm quickly converges to one of many inequivalent unstable solutions. When the policy update 2 is computed using *all hands*, CFR converges to a solution that exhibits zero collaboration. The expected rewards are in Table 1.

A salient detail is that the inter-team transfer is $\Delta < 0.01$ chip per hand, while the average inter-team transfer is $\delta_{ij} = 0.25$ chip per hand. This shows that CFR has learned a non-collaborative strategy: players gain chips by defecting from partners, rather than collaborating with them.

Collective Reward Engineering Using CFR with tied rewards (`CFR-tied-`r) improves performance over `CFR`: the learned strategy has nontrivial amounts of chips going from one team to another ($\Delta > 0$). It also has fewer chips going from one person to their partner (δ_{ij} is smaller), indicating a higher level of collaboration. This suggests that engineering rewards to be tied for players in the same team improves the strategy, as well as facilitate cooperation.

We observed that under the strategy learned using `CFR-tied-`r, the first player checks only if they hold a Queen, and their partner plays very aggressively if they also hold a Queen. This suggests that the two players have learned to communicate as part of their collaboration.

State Inference Model The players with access to an oracle (`CFR+oracle`) had the smallest intra-team transfer δ, and the highest Δ, meaning they displayed the most collaboration. In competition, `CFR+oracle` has modest victories against `CFR` and `CFR-tied-`r, suggesting that using a (learned) state inference model can improve the strategy.

However, adding a realistic noise model completely erased all effectiveness: `CFR+oracle-`p is always dominated by the non-oracle models. This suggests that a learned state inference model would need $\geq 90\%$ accuracy, and be able to infer a partner's hand after at most 4 actions in this setting.

4 Future Work

We aim to investigate how effective a state inference model can be made when learning using CFR. Our experiments suggest that any learned state inference model needs high levels of accuracy and inference speed during a round, in order to enable standard CFR to find stronger strategies. Hence, it would be interesting to investigate how to *learn* an accurate state inference model and strategy jointly using (extensions of) CFR. Moreover, an open question is how state inference could enable complex strategic behavior, such as bluffing and implicit communication.

References

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in neural information processing systems*, pages 1729–1736, 2008.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in no-limit poker. *arXiv preprint arXiv:1701.01724*, 2017.

J. N. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch. Learning with Opponent-Learning Awareness. *ArXiv e-prints*, September 2017.

H. W. Kuhn. Simplified two-person poker. *Contributions to the Theory of Games*, 1950.

Samuel Barrett, Peter Stone, Sarit Kraus, and Avi Rosenfeld. Teamwork with limited knowledge of teammates. 2013.