

# Combining Deep Reinforcement Learning and Search for Imperfect-Information Games

Noam Brown\* Anton Bakhtin\* Adam Lerer Qucheng Gong  
Facebook AI Research  
{noambrown,yolo,alerer,qucheng}@fb.com

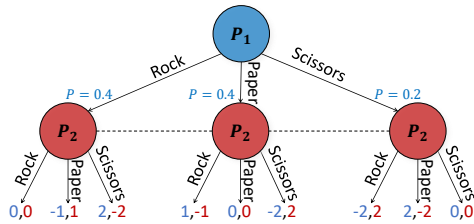
## Abstract

The combination of deep reinforcement learning and search at both training and test time is a powerful paradigm that has led to a number of successes in single-agent settings and perfect-information games, best exemplified by the success of AlphaZero. However, algorithms of this form have been unable to cope with imperfect-information games. This paper presents ReBeL, a general framework for self-play reinforcement learning and search for imperfect-information games. In the simpler setting of perfect-information games, ReBeL reduces to an algorithm similar to AlphaZero. Results show ReBeL leads to low exploitability in benchmark imperfect-information games and achieves superhuman performance in heads-up no-limit Texas hold'em poker, while using far less domain knowledge than any prior poker AI. We also prove that ReBeL converges to a Nash equilibrium in two-player zero-sum games in tabular settings.

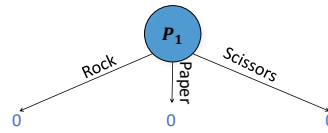
## 1 Introduction

Combining reinforcement learning with search at both training and test time (**RL+Search**) has led to a number of major successes in AI in recent years. For example, the AlphaZero algorithm achieves state-of-the-art performance in the perfect-information games of Go, chess, and shogi [55].

However, existing RL+Search algorithms do not work in imperfect-information games because they make a number of assumptions that no longer hold in these settings. An example of this is illustrated in Figure 1a, which shows a modified form of Rock-Paper-Scissors in which the winner receives two points (and the loser loses two points) when either player chooses Scissors [16]. The figure shows the game in a sequential form in which player 2 acts after player 1 but does not observe player 1's action.



(1a) Variant of Rock-Paper-Scissors in which the optimal  $P_1$  policy is (R=0.4, P=0.4, S=0.2). Terminal values are color-coded. The dotted lines mean  $P_2$  does not know which node they are in.



(1b) The  $P_1$  subgame when using perfect-information one-ply lookahead search. Leaf values are determined by the full-game equilibrium. There is insufficient information for finding (R=0.4, P=0.4, S=0.2).

The optimal policy for both players in this modified version of the game is to choose Rock and Paper with 40% probability, and Scissors with 20%. In that case, each action results in an expected value of zero. However, as shown in Figure 1b, if player 1 were to conduct one-ply lookahead search as is

\*Equal contribution

done in perfect-information games (in which the equilibrium value of a state is substituted at a leaf node), then there would not be enough information for player 1 to arrive at this optimal policy.

This illustrates a critical challenge of imperfect-information games: unlike perfect-information games and single-agent settings, the value of an action may depend on the probability it is chosen, and more generally may depend on the entire policy. Thus, existing RL+Search algorithms such as AlphaZero are not sound in imperfect-information games. Recent AI breakthroughs in imperfect-information games have highlighted the importance of search at test time [42, 13, 15, 39], but combining RL and search during training in imperfect-information games has been an open problem. Other recent imperfect-information game successes have relied on RL alone without leveraging search [4, 61].

This paper introduces ReBeL (Recursive Belief-based Learning), a general RL+Search algorithm that converges to a Nash equilibrium in two-player zero-sum games. Our method builds on prior work in which the notion of “state” is expanded to include the probabilistic belief distribution of all agents about what state they may be in, given the available common knowledge information and the policies of all agents. Our algorithm trains a value network and a policy network for these states through self-play reinforcement learning. Additionally, the algorithm uses the value and policy network for search during self play.

Our goal in this paper is not to chase state-of-the-art performance by any means necessary. Instead, our goal is to develop a simple, flexible, effective algorithm that leverages as little expert domain knowledge as possible. Experimental results show that despite its simplicity, ReBeL is effective in large-scale two-player zero-sum imperfect-information games and defeats a top human professional with statistical significance in the benchmark game of heads-up no-limit Texas hold’em poker while using far less expert domain knowledge than any previous poker AI. In perfect-information games, ReBeL simplifies to an algorithm similar to AlphaZero, with the major differences being that search occurs over a fixed depth, the search policy is played for the entirety of the subgame rather than just the next action, and the value network is trained using bootstrapping.

## 2 Related Work

At a high level, our framework resembles past RL+Search algorithms used in perfect-information games [59, 56, 1, 55, 51]. These algorithms train a value network through self play. During training, a search algorithm is used in which the values of leaf nodes are determined via the value function. Additionally, a policy network may be used to guide search. These forms of RL+Search have been critical to achieving superhuman performance in benchmark perfect-information games. For example, so far no AI agent has achieved superhuman performance in Go without using search at both training and test time. However, these RL+Search algorithms are not theoretically sound in imperfect-information games and have not been shown to be successful in such settings.

A critical element of our imperfect-information RL+Search framework is to use an expanded notion of “state”, which we refer to as a **public belief state (PBS)**. PBSs are defined by a common-knowledge belief distribution over states, determined by the public observations shared by all agents and the known policies of all agents. PBSs can be viewed as a multi-agent generalization of belief states used in partially observable Markov decision processes (POMDPs) [33]. The concept of PBSs originated in work on decentralized multi-agent POMDPs [45, 47, 20] and has been widely used since then in imperfect-information games more broadly [42, 21, 53, 31].

Our work most closely builds upon the poker AI DeepStack [42], in which a value function for PBSs is used for search at test time. However, DeepStack’s value function was trained not through self-play RL, but rather by generating random PBSs, including random probability distributions, and estimating their values using search. This would be like learning a value function for Go by randomly placing stones on the board. This is not an efficient way of learning a value function because the vast majority of randomly generated situations would not be relevant in actual play. DeepStack coped with this by using handcrafted features to reduce the dimensionality of the public belief state space, by sampling PBSs from a distribution based on expert domain knowledge, and by using domain-specific abstractions to circumvent the need for a value network when close to the end of the game.

An alternative approach for depth-limited search in imperfect-information games that does not use a value function for PBSs was used in the Pluribus poker AI to defeat elite human professionals in multiplayer poker for the first time [16, 15]. This approach trains a population of “blueprint” policies without using search. At test time, the approach conducts depth-limited search by allowing each agent to choose a blueprint policy from the population at leaf nodes. The value of the leaf node is the

expected value of each agent playing their chosen blueprint policy against all the other agents' choice for the rest of the game. While this approach has been successful in poker, it does not use search during training and therefore requires strong blueprint policies to be computed without search. Also, the computational cost of the search algorithm is linear with the number of blueprint policies.

### 3 Notation and Background

We assume throughout this work that the rules of the game, as well as any models and algorithms used by our agent, are common knowledge. However, the outcome of stochastic algorithms (i.e., the random seeds) are not known.

Our notation is based on that of factored observation games [36] which is a modification of partially observable stochastic games [27] that distinguishes between private and public observations. We consider a game with  $\mathcal{N} = \{1, 2, \dots, N\}$  agents. We provide theoretical and empirical results only for when  $|\mathcal{N}| = 2$ , though related techniques have been shown to be successful in practice in certain settings with more agents [15].

A **world state**  $w \in \mathcal{W}$  is a state in the game.  $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \dots \times \mathcal{A}_N$  is the space of joint actions.  $\mathcal{A}_i(w)$  denotes the legal actions for agent  $i$  at  $w$  and  $a = (a_1, a_2, \dots, a_N) \in \mathcal{A}$  denotes a joint action. After a joint action  $a$  is chosen, a transition function  $\mathcal{T}$  determines the next world state  $w'$  drawn from the probability distribution  $\mathcal{T}(w, a) \in \Delta\mathcal{W}$ . After joint action  $a$ , agent  $i$  receives a reward  $\mathcal{R}_i(w, a)$ .

Upon transition from world state  $w$  to  $w'$  via joint action  $a$ , agent  $i$  receives a **private observation** from a function  $\mathcal{O}_{\text{priv}(i)}(w, a, w')$ . Additionally, all agents receive a **public observation** from a function  $\mathcal{O}_{\text{pub}}(w, a, w')$ . Public observations may include observations of publicly taken actions by agents. For example, in many recreational games, including poker, all betting actions are public.

A **history** (also called a trajectory) is a finite sequence of legal actions and world states, denoted  $h = (w^0, a^0, w^1, a^1, \dots, w^t)$ . An **infostate** (also called an **action-observation history (AOH)**) for agent  $i$  is a sequence of an agent's observations and actions  $s_i = (O_i^0, a_i^0, O_i^1, a_i^1, \dots, O_i^t)$  where  $O_i^k = (\mathcal{O}_{\text{priv}(i)}(w^{k-1}, a^{k-1}, w^k), \mathcal{O}_{\text{pub}}(w^{k-1}, a^{k-1}, w^k))$ . The unique infostate corresponding to a history  $h$  for agent  $i$  is denoted  $s_i(h)$ . The set of histories that correspond to  $s_i$  is denoted  $\mathcal{H}(s_i)$ .

A **public state** is a sequence  $s_{\text{pub}} = (O_{\text{pub}}^0, O_{\text{pub}}^1, \dots, O_{\text{pub}}^t)$  of public observations. The unique public state corresponding to a history  $h$  and an infostate  $s_i$  is denoted  $s_{\text{pub}}(h)$  and  $s_{\text{pub}}(s_i)$ , respectively. The set of histories that match the sequence of public observation of  $s_{\text{pub}}$  is denoted  $\mathcal{H}(s_{\text{pub}})$ .

As an example, consider a game in which two players roll two six-sided dice each. One die of each player is publicly visible, but the other die is only observed by the player who rolled it. Suppose player 1 rolls a 3 and a 4 (with 3 being the hidden die), and player 2 rolls a 5 and a 6 (with 5 being the hidden die). The history is  $h = ((3, 4), (5, 6))$ . The set of histories corresponding to player 2's infostate is  $\mathcal{H}(s_2) = \{((x, 4), (5, 6)) \mid x \in \{1, 2, 3, 4, 5, 6\}\}$ , so  $|\mathcal{H}(s_2)| = 6$ . The set of histories corresponding to  $s_{\text{pub}}$  is  $\mathcal{H}(s_{\text{pub}}) = \{((x, 4), (y, 6)) \mid x, y \in \{1, 2, 3, 4, 5, 6\}\}$ , so  $|\mathcal{H}(s_{\text{pub}})| = 36$ .

Public states provide an easy way to reason about common knowledge in a game. All agents observe the same public sequence  $s_{\text{pub}}$ , and therefore it is common knowledge among all agents that the true history is some  $h \in \mathcal{H}(s_{\text{pub}})$ .<sup>2</sup>

An agent's **policy**  $\pi_i$  is a function mapping from an infostate to a probability distribution over actions. A **policy profile**  $\pi$  is a tuple  $(\pi_1, \pi_2, \dots, \pi_N)$ . We also define a policy for a history  $h$  as  $\pi_i(h) = \pi_i(s_i(h))$  and  $\pi(h) = (\pi_1(s_1(h)), \pi_2(s_2(h)), \dots, \pi_N(s_N(h)))$ . The expected sum of future rewards (also called the **expected value (EV)**) for agent  $i$  in history  $h$  when all agents play policy profile  $\pi$  is denoted  $v_i^\pi(h)$ . The EV for the entire game is denoted  $v_i(\pi)$ . A **Nash equilibrium** is a policy profile such that no agent can achieve a higher EV by switching to a different policy [44]. Formally,  $\pi^*$  is a Nash equilibrium if for every agent  $i$ ,  $v_i(\pi^*) = \max_{\pi_i} v_i(\pi_i, \pi_{-i}^*)$  where  $\pi_{-i}$  denotes the policy of all agents other than  $i$ . A **Nash equilibrium policy** is a policy  $\pi_i^*$  that is part of some Nash equilibrium  $\pi^*$ .

<sup>2</sup>As explained in [36], it may be possible for agents to infer common knowledge beyond just public observations. However, doing this additional reasoning is inefficient both theoretically and practically.

## 4 From World States to Public Belief States

A **perfect-information subgame** is defined by a root world state  $w$  and all states that can be reached from that point forward. In other words, it is identical to the original game except it starts at  $w$ . In two-player zero-sum perfect-information games, every world state  $w$  has a unique value  $v_i(w)$  for each agent  $i$  defined by both agents playing a Nash equilibrium  $\pi^*$  in the subgame rooted at that world state. While there might be multiple Nash equilibria in a subgame, they all result in the same EV for the agents in a two-player zero-sum game. A typical goal of RL algorithms for perfect-information games is to learn the value function  $V_i : \mathcal{W} \rightarrow \mathbb{R}$  for each agent  $i$  mapping a world state to its value in an equilibrium. With such a value function, one can find the optimal policy for a world state by solving a depth-limited subgame that extends only for a limited number of actions in the future, where the values of leaf states (that is, states that have legal actions in the full game but not in the depth-limited subgame) are set according to the value function [54, 50].

However, world states and histories do not necessarily have unique values in imperfect-information games because their values depend not just on the true state of the world, but also on what each agent knows, what each agent knows about what the other agents know, etc. In other words, it depends on the **common knowledge** [2] among the agents.

**Public belief states** (PBSs) generalize the notion of “state value” to imperfect-information games. A PBS  $\beta$  is a common-knowledge probability distribution over histories in  $\mathcal{H}(s_{\text{pub}})$  for some public state  $s_{\text{pub}}$ . In perfect-information games, PBSs reduce to histories, which in two-player zero-sum games effectively reduce to world states.

Any PBS that can arise in play (i.e., that can arise from the agents playing some policy profile  $\pi$ ) can always be described by a joint probability distribution over the agents’ possible infostates, given some public state  $s_{\text{pub}}$  [47, 52]. That is, the probability of any history in the PBS can be computed when given the joint probability distribution over infostates and the rules of the game. For example, in poker a PBS consists of the player actions and public cards, along with a probability distribution over possible hands for each player. Formally, let  $S_i(s_{\text{pub}})$  be the set of infostates that player  $i$  may be in given a public state  $s_{\text{pub}}$ . Then  $\text{PBS } \beta = (s_{\text{pub}}, \Delta S_1(s_{\text{pub}}), \dots, \Delta S_N(s_{\text{pub}}))$ .<sup>3</sup>

We can generalize the notion of “state value” to imperfect-information games by defining a **subgame** to be rooted at a PBS.<sup>4</sup> One interpretation of such a subgame is that at the start of the subgame a history is sampled based on the joint probability distribution of the PBS, and then the game proceeds as it would in the original game.

Just as in perfect-information subgames, the optimal agent policies in a subgame rooted at a PBS do not depend on anything (policies, observations, etc.) that came before the PBS. Thus, a two-player zero-sum subgame rooted at a PBS  $\beta$  has a unique value  $V_1(\beta) = \sum_{h \in \mathcal{H}(s_{\text{pub}}(\beta))} (p(h)v_1^{\pi^*}(h))$ , where  $s_{\text{pub}}(\beta)$  is the public state corresponding to  $\beta$  and  $\pi^*$  is a Nash equilibrium in the subgame. Since the game is zero-sum,  $V_1(\beta) = -V_2(\beta)$ .

Just as one can compute an optimal policy in perfect-information games via search by learning a value function for world states, we show that one can compute an optimal policy in imperfect-information games via search by learning a value function  $V_1 : \mathcal{B} \rightarrow \mathbb{R}$ , where  $\mathcal{B}$  is the continuous space of PBSs. Existing depth-limited search algorithms for imperfect-information games require the EVs of *infostates* for PBSs, rather than the value of the PBS as a whole [17, 42]. The EV of infostate  $s_i$  in  $\beta$  assuming  $\pi^*$  is played is

$$v_i^{\pi^*}(s_i|\beta) = \sum_{h \in \mathcal{H}(s_i)} p(h|s_i, \beta) v_i^{\pi^*}(h)$$

Theorem 1 proves that infostate EVs for  $\beta$  under some Nash equilibrium  $\pi^*$  can be derived from  $V_i$ .

<sup>3</sup>Frequently, a PBS can be summarized even more compactly by discarding parts of the history that are no longer strategically relevant. For example, in poker we do not need to track the entire history of actions when given a probability distribution over each players’ hands. We just need to track the amount of money each player has in the pot, the public board cards, and whether there were any bets in the current round.

<sup>4</sup>Past work defines a subgame to be rooted at a public state [17, 9, 43, 42, 12, 35, 36, 57, 52]. We break with this definition because imperfect-information subgames rooted at a public state do not have well-defined values.

**Theorem 1.** Let  $\tilde{V}_i$  be the extension of  $V_i$  to unnormalized belief distributions. For any (normalized) belief  $\beta$  and any subgradient  $-\bar{g}_i$  of  $-\tilde{V}_i(\beta)$  with respect to  $\beta$ ,  $V_i(\beta) + \bar{g}_i \cdot \hat{s}_i = v_i^{\pi^*}(s_i|\beta)$  for some Nash equilibrium policy  $\pi^*$ , where  $\hat{s}_i$  is the unit vector in direction  $s_i$ .

All proofs are presented in the appendix.

Thus, in theory, learning  $V_1$  is sufficient for using the search algorithms in this paper. Nevertheless, in practice we learn an infostate-value function  $\hat{v} : \mathcal{B} \rightarrow \mathbb{R}^{|S_1|+|S_2|}$  that directly approximates for each  $s_i$  the average of the sampled  $v_i^{\pi^*}(s_i|\beta)$  values produced by our RL+Search procedure at  $\beta$ .

Unlike the total PBS value  $V_1(\beta)$ , the infostate values may depend on which Nash equilibrium is played in the subgame. Each execution  $k$  of our RL+Search algorithm may converge to a different  $\pi^{*,k}$  in  $\beta$ , so the samples of  $v_i^{\pi^{*,k}}(\beta)$  may not be identical. Nevertheless, Theorem 2 states that the average of valid samples of  $v_i^{\pi^{*,k}}(\beta)$  corresponds to  $v_i^{\pi^*}(\beta)$  for some other minimax policy  $\pi^*$ . Therefore  $\hat{v}$  should approximate  $v_i^{\pi^*}$  for some minimax policy.

**Theorem 2.** Let  $X$  be the vector of infostate EVs in PBS  $\beta$  corresponding to minimax policy profile  $\pi^{*,X}$ , and let  $Y$  be the vector of infostate EVs in  $\beta$  corresponding to minimax policy profile  $\pi^{*,Y}$ . Then  $\lambda X + (1 - \lambda)Y$  is the vector of infostate EVs in  $\beta$  corresponding to minimax policy profile  $\lambda\pi^{*,X} + (1 - \lambda)\pi^{*,Y}$  for  $0 \leq \lambda \leq 1$ .

Given that agents play according to policy profile  $\pi$ , the PBS that arises at public state  $s_{\text{pub}}$ , the infostate  $s_i$ , and the history  $h$  is denoted  $\beta_{s_{\text{pub}}}^\pi$ ,  $\beta_{s_i}^\pi$ , and  $\beta_h^\pi$ , respectively.

A **depth-limited subgame** is a subgame (again rooted at a PBS) that extends only for some limited number of actions into the future. In this paper, search is performed over a fixed-size depth-limited subgame (as opposed to Monte Carlo Tree Search, which grows the subgame as more search iterations are performed [24]), and we assume that all histories sharing a public state are either all in the subgame or all not in the subgame. A history  $z$  that has children in the full game but does not have children in the subgame is a **leaf node**, and agent  $i$  receives a reward of  $\hat{v}(s_i(z)|\beta_z^\pi)$  at such a history, where  $\pi$  is the policy profile in the subgame. This means that the value of a leaf node is conditional on the beliefs at that leaf node, which in turn are conditional on the policy in the subgame.

## 5 Self Play Reinforcement Learning and Search for Public Belief States

At a high level, ReBeL, shown in Algorithm 1, is similar to RL+Search algorithms used for perfect-information games, but operating on PBSs rather than world states. At the start of the game, a depth-limited subgame rooted at the initial PBS  $\beta_r$  is generated. This subgame is solved (i.e., a Nash equilibrium is approximated) by running  $T$  iterations of an iterative equilibrium-finding algorithm and using the learned value network  $\hat{v}$  to approximate leaf values on every iteration. The infostate values at  $\beta_r$  are added as training examples for  $\hat{v}$  and (optionally) the policies in the subgame are added as training examples for the policy network. Finally, a leaf node  $z$  is sampled and the process repeats with the PBS at  $z$  being the new subgame root. Detailed pseudocode is provided in Appendix B.

Section 5.1 explains how to solve a depth-limited subgame, given a value network. Section 5.2 covers in detail our self-play reinforcement learning algorithm that uses the techniques in Section 5.1 for search. Section 5.3 describes how to improve these techniques by adding a policy network.

### 5.1 Search in a depth-limited imperfect-information subgame

There exist a number of iterative algorithms for solving imperfect-information games [6, 62, 30, 38, 37]. Our framework is flexible with respect to the choice of a search algorithm.

We assume that the search algorithm used is an iterative self-play algorithm. On each iteration  $t$ , the algorithm determines a policy profile  $\pi^t$ . Next, the value of every leaf node  $z$  is set according to  $\hat{v}(s_i(z)|\beta_z^{\pi^t})$  or  $\hat{v}(s_i(z)|\beta_z^{\bar{\pi}^t})$ , depending on the algorithm, where  $\bar{\pi}^t$  denotes the average policy profile over iterations 1 to  $t$ . Given  $\pi^t$  and the leaf node values, each infostate in  $\beta_r$  has a well-defined value. This vector of values, denoted  $v^{\pi^t}(\beta_r)$ , is computed and stored. Next, the algorithm chooses a new policy profile  $\pi^{t+1}$ , and the process repeats for  $T$  iterations. For many algorithms, including **Counterfactual Regret Minimization (CFR)** [62, 17, 42], the average policy profile  $\bar{\pi}^T$  converges to a Nash equilibrium as  $T \rightarrow \infty$ .

---

**Algorithm 1** ReBeL: RL and Search for Imperfect-Information Games
 

---

```

function SELFPLAY( $\beta_r, \theta^v, \theta^\pi, D^v, D^\pi$ ) ▷  $\beta_r$  is the current PBS
  while !IS TERMINAL( $\beta_r$ ) do
     $G \leftarrow \text{CONSTRUCTSUBGAME}(\beta_r)$ 
     $\bar{\pi}, \pi^{t_{\text{warm}}} \leftarrow \text{INITIALIZEPOLICY}(G, \theta^\pi)$  ▷  $t_{\text{warm}} = 0$  and  $\pi^0$  is uniform if no warm start
     $G \leftarrow \text{SETLEAFVALUES}(G, \bar{\pi}, \pi^{t_{\text{warm}}}, \theta^v)$ 
     $v(\beta_r) \leftarrow \text{COMPUTE EV}(G, \pi^{t_{\text{warm}}})$ 
     $t_{\text{sample}} \sim \text{unif}\{t_{\text{warm}} + 1, T\}$  ▷ Sample an iteration
    for  $t = (t_{\text{warm}} + 1) \dots T$  do
       $\pi^t \leftarrow \text{UPDATEPOLICY}(G, \pi^{t-1})$ 
       $\bar{\pi} \leftarrow \frac{t-1}{t} \bar{\pi} + \frac{1}{t} \pi^t$ 
       $G \leftarrow \text{SETLEAFVALUES}(G, \bar{\pi}, \pi^t, \theta^v)$ 
       $v(\beta_r) \leftarrow \frac{t-1}{t} v(\beta_r) + \frac{1}{t} \text{COMPUTE EV}(G, \pi^t)$ 
      if  $t = t_{\text{sample}}$  then
         $\beta'_r \leftarrow \text{SAMPLELEAF}(G, \pi^t)$  ▷ Sample a leaf PBS according to the new policies
      Add  $\{\beta_r, v(\beta_r)\}$  to  $D^v$  ▷ Add to value net training data
      for  $\beta \in G$  do ▷ Loop over the PBS at every public state in  $G$ 
        Add  $\{\beta, \bar{\pi}(\beta)\}$  to  $D^\pi$  ▷ Add to policy net training data (optional)
       $\beta_r \leftarrow \beta'_r$ 

```

---

After solving a subgame rooted at PBS  $\beta_r$  with an iterative algorithm that has run for  $T$  iterations, the value vector  $(\sum_{t=1}^T v^{\pi^t}(\beta_r))/T$  is added to the training data for  $\hat{v}(\beta_r)$ .<sup>5</sup>

Prior work on search in imperfect-information games has used the **CFR Decomposition (CFR-D)** algorithm [17, 42]. Appendix E introduces **CFR-AVE**, a modification of CFR-D that sets the value of a leaf node  $z$  based on  $\bar{\pi}^t$  rather than  $\pi^t$ , which addresses some weaknesses of CFR-D. Section 8 also shows experimental results for **fictitious play (FP)** [6].

## 5.2 Self-play reinforcement learning

Algorithm 1 learns values for PBSs through self play. After solving a subgame rooted at PBS  $\beta_r$ , the value vector for the root infostates is added to the training dataset for  $\hat{v}$ . Next, a leaf PBS  $\beta'_r$  is sampled and a new subgame rooted at  $\beta'_r$  is solved. This process repeats until the game ends.

Since the subgames are solved using an iterative algorithm, we want  $\hat{v}$  to be accurate for leaf PBSs on every iteration. Therefore, a leaf node  $z$  is sampled according to  $\pi^t$  on a random iteration  $t \sim \text{unif}\{0, T-1\}$ , where  $T$  is the number of iterations of the search algorithm.<sup>6</sup> To ensure sufficient exploration, one agent samples random actions with probability  $\epsilon > 0$ .<sup>7</sup> In CFR-D  $\beta'_r = \beta_z^{\pi^t}$ , while in CFR-AVE and FP  $\beta'_r = \beta_z^{\bar{\pi}^t}$ .

Eventually, a subgame rooted at  $\beta_r^*$  is reached near the end of the game that does not contain leaf nodes (i.e., the subgame is not depth-limited).  $\hat{v}$  will therefore learn correct values  $v^{\pi^*}(s_i | \beta_r^*)$  for every root infostate  $s_i$  and for some Nash equilibrium  $\pi^*$  (except for an error term that disappears as  $T \rightarrow \infty$ ). In the future, when  $\beta_r^*$  is a leaf PBS of a different subgame, it will be possible to more accurately compute the value of that subgame. In this way, accurate PBS values will “bubble up” the game tree and  $\hat{v}$  will increase in accuracy over time.

Theorem 3 states that, with perfect function approximation, running Algorithm 1 will produce a value network whose error is bounded by  $\mathcal{O}(\frac{1}{\sqrt{T}})$  after a finite amount of time for any PBS that could be encountered during play, where  $T$  is the number of CFR iterations being run in subgames.

**Theorem 3.** *Consider an idealized value approximator that returns the most recent sample of the value for sampled PBSs, and 0 otherwise. Running Algorithm 1 with  $T$  iterations of CFR in each subgame will, after a finite amount of time, produce a value approximator that has error of at most  $\frac{C}{\sqrt{T}}$  for any PBS that could be encountered during play, where  $C$  is a game-dependent constant.*

---

<sup>5</sup>For some algorithms, including CFR-AVE and FP, an alternative is to add the value vector  $v^{\pi^T}(\beta_r)$ .

<sup>6</sup>For FP, we pick a random agent  $i$  and sample according to  $(\pi_i^t, \bar{\pi}_{-i}^t)$  to reflect the search operation.

<sup>7</sup>The algorithm is still correct if all agents sample random actions with probability  $\epsilon$ , but that is less efficient because the value of a leaf node that can only be reached if both agents go off policy is irrelevant.

### 5.3 Adding a policy network

Algorithm 1 will result in  $\hat{v}$  converging correctly even if a policy network is not used. However, if a policy network is not used then in the first several iterations  $t$  of running a search algorithm in the subgames,  $\bar{\pi}^t$  may be very far from an equilibrium. Since, in some search algorithms,  $\bar{\pi}^t$  determines the PBSs at the leaf nodes of a subgame, not using a policy network means the learned value network must be accurate over a wide domain of PBSs. Additionally, adding an accurate policy network will reduce the number of search iterations necessary to closely approximate a Nash equilibrium.

Algorithm 1 can train a policy network  $\hat{\Pi} : \beta \rightarrow (\Delta \mathcal{A})^{|S_1|+|S_2|}$  by adding  $\bar{\pi}^T(\beta)$  for each PBS  $\beta$  in the subgame to a training dataset each time a subgame is solved (i.e.,  $T$  iterations of CFR have been run in the subgame). Appendix J describes our warm start technique, which is based on [11].

### 5.4 Algorithm behavior in perfect-information games

Perfect-information games can be viewed as a special case of imperfect-information games in which public states are equivalent to histories, and therefore have the same value as world states. Since the value of a leaf node in a perfect-information subgame does not depend on the policy in the subgame, only one search iteration is required to solve a subgame.

Thus, in perfect-information games Algorithm 1 reduces to an algorithm similar to AlphaZero. The major differences are that AlphaZero plays just a single action before solving a new subgame while ReBeL plays the subgame policy until reaching a leaf node, and AlphaZero grows the size of the subgame during search, and AlphaZero trains on the final reward received at the end of the game.

## 6 Playing According to an Equilibrium at Test Time

In perfect-information games, it is fairly trivial to compute an optimal policy once an exact value function  $\hat{v}$  is computed. This is not the case in imperfect-information games. As an example of the problem we seek to address, consider again the game of modified Rock-Paper-Scissors illustrated in Figure 1a. Suppose that  $\hat{v}$  is perfect and we are now playing against a real opponent where we are player 2. Suppose that player 1 has just acted, In order to now conduct search as player 2, our algorithm requires a belief distribution over states. What should this belief distribution be?

An intuitive choice, which is referred to as **unsafe** search [26, 23], is to first run CFR for  $T$  iterations for player 1’s first move (for some large  $T$ ) and eventually arrive at a player 1 policy such as  $(R = 0.41, P = 0.39, S = 0.2)$ . Unsafe search passes down the beliefs resulting from that policy, and then computes our optimal policy as player 2. This would result in a policy of  $(R = 0, P = 1, S = 0)$  for player 2. Clearly, this is not a Nash equilibrium. Moreover, if our opponent knew we would end up playing this policy (which we assume they would know since we assume they know the algorithm we run to generate the policy), then they could counter us by playing  $(R = 0, P = 0, S = 1)$ .

This problem demonstrates the need for **safe** search, which is a search algorithm that ensures we play a Nash equilibrium policy in expectation. Importantly, it is *not* necessary for the policy that the algorithm outputs to always be a Nash equilibrium. It is only necessary that the algorithm outputs a Nash equilibrium policy *in expectation*. For example, in modified Rock-Paper-Scissors it is fine for an algorithm to output a policy of 100% Rock, so long as the probability it outputs that policy is 40%.

All past approaches for conducting safe search introduce additional constraints to the search algorithm to ensure the value of each infostate in the root PBS matches the value from  $\hat{v}$  [17, 43, 12, 57]. Those additional constraints ensures search approximates a Nash equilibrium if  $\hat{v}$  is accurate, but the constraints also hurt performance in practice compared to unsafe search [17, 12] and greatly complicate search, so they were never fully used in any competitive agent. Instead, all previous search-based agents used unsafe search either partially or entirely [42, 13, 16, 15, 53].

Moreover, using a different algorithm for search at test time than was used for search during training may result in encountering PBSs at test time that were not encountered during training and therefore result in poor approximations for the value and policy network in a self-play agent.

We now prove that safe search can be achieved without any additional constraints by simply running the exact same algorithm at test time that was used during training. Specifically, when conducting search at test time we pick a random iteration and assume all agents play according to that iteration’s policy profile for the entire subgame. This leads to a PBS leaf node, which defines a new subgame, and the process repeats. Since the opponent does not know which iteration we selected, they are not able to exploit our assumption about the belief distribution. Theorem 4, the proof of which is in

Section I, states that this algorithm approximates a Nash equilibrium. Specifically, Theorem 4 states that once a value network is trained according to Theorem 3, using Algorithm 1 at test time (without off-policy exploration) will approximate a Nash equilibrium.

**Theorem 4.** *If Algorithm 1 is run at test time with no off-policy exploration, a value network with error at most  $\delta$  for any leaf PBS that was trained to convergence as described in Theorem 3, and with  $T$  iterations of CFR being used to solve subgames, then the algorithm plays a  $(\delta C_1 + \frac{\delta C_2}{\sqrt{T}})$ -Nash equilibrium, where  $C_1, C_2$  are game-specific constants.*

In other words, the same algorithm we describe for training also approximates a Nash equilibrium at test time. This result applies regardless of how the value network was trained and therefore can be applied to prior algorithms that use PBS value functions [42, 53].

Since a random iteration is selected, there is a risk that we may select a very early iteration, or even the first iteration, in which the policy is extremely poor. This can be mitigated by using modern equilibrium-finding algorithms, such as Linear CFR [14], that assign little or no weight to the early iterations that are played.

## 7 Experimental Setup

We measure **exploitability** of a policy  $\pi^*$ , which is  $\sum_{i \in \mathcal{N}} \max_{\pi} v_i(\pi, \pi^*_i) / |\mathcal{N}|$ . All CFR experiments use alternating-updates Linear CFR [14]. All FP experiments use alternating-updates Linear Optimistic FP, which is a novel variant we present in Appendix D.

We evaluate on the benchmark imperfect-information games of heads-up no-limit Texas hold’em poker (HUNL) and Liar’s Dice. The rules for both games are provided in Appendix C. We also evaluate our techniques on turn endgame hold’em (TEH), a variant of no-limit Texas hold’em in which both players automatically check/call for the first two of the four betting rounds in the game.

In HUNL and TEH, we reduce the action space to at most nine actions using domain knowledge of typical bet sizes. However, our agent responds to any “off-tree” action at test time by adding the action to the subgame [16, 15]. The bet sizes and stack sizes are randomized during training. For TEH we train on the full game and measure exploitability on the case of both players having \$20,000, unperturbed bet sizes, and the first four board cards being  $3\spadesuit 7\heartsuit T\diamond K\spadesuit$ .

For HUNL, our agent uses far less domain knowledge than any prior competitive AI agent. Additionally, our AI agent is trained on all stack sizes between 5,000 and 25,000 chips, rather than just the standard 20,000. Appendix F discusses the poker domain knowledge we leveraged in ReBeL.

We approximate the value and policy functions using artificial neural networks. Both networks are MLPs with GeLU [29] activation functions and LayerNorm [3]. Both networks are trained with Adam [34]. We use pointwise Huber loss as the criterion for the value function and mean squared error (MSE) over probabilities for the policy. In preliminary experiments we found MSE for the value network and cross entropy for the policy network did worse. See Appendix G for the hyperparameters.

We use PyTorch [48] to train the networks. We found data generation to be the bottleneck due to the sequential nature of the FP and CFR algorithms and the evaluation of all leaf nodes on each iteration. For this reason we use a single machine for training and up to 128 machines with 8 GPUs each for data generation. We release an implementation<sup>8</sup> for Liar’s Dice to showcase the algorithms.

## 8 Experimental Results

Figure 2 shows ReBeL, using a learned value network, reaches a level of exploitability in TEH equivalent to running about 125 iterations of full-game tabular CFR. For context, top poker agents typically use between 100 and 1,000 tabular CFR iterations [5, 42, 13, 16, 15].

Table 1 shows results for ReBeL in HUNL. We compare ReBeL to BabyTartanian8 [10] and Slumbot, prior champions of the Computer Poker Competition, and the local best response (LBR) [41] algorithm. We also present results against Dong Kim, a top human HUNL expert that did best among the four top humans that played against Libratus. Kim played 7,500 hands. Variance was reduced by using AIVAT [18]. ReBeL played faster than 2 seconds per hand and never needed more than 5 seconds for a decision.

<sup>8</sup><https://github.com/facebookresearch/rebel>



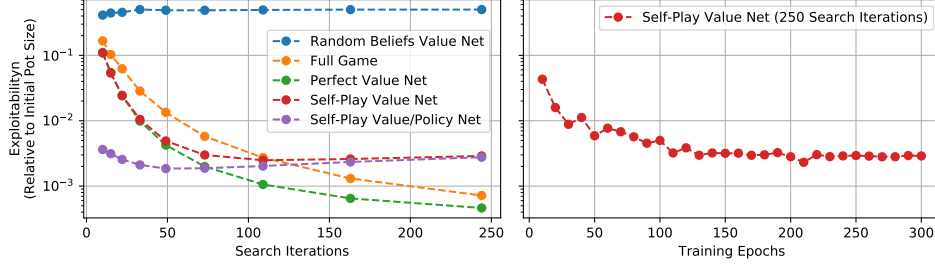


Figure 2: Convergence of different techniques in TEH. All subgames are solved using CFR-AVE. Perfect Value Net uses an oracle function to return the exact value of leaf nodes on each iteration. Self-Play Value Net uses a value function trained through self play. Self-Play Value/Policy Net additionally uses a policy network to warm start CFR. Random Beliefs trains the value net by sampling PBSs at random.

Bot Name	Slumbot	BabyTartanian8 [10]	LBR [41]	Top Humans
DeepStack [42]	-	-	$383 \pm 112$	-
Libratus [13]	-	$63 \pm 14$	-	$147 \pm 39$
Modicum [16]	$11 \pm 5$	$6 \pm 3$	-	-
ReBeL ( <i>Ours</i> )	$45 \pm 5$	$9 \pm 4$	$881 \pm 94$	$165 \pm 69$

Table 1: Head-to-head results of our agent against benchmark bots BabyTartanian8 and Slumbot, as well as top human expert Dong Kim, measured in thousandths of a big blind per game. We also show performance against LBR [41] where the LBR agent must call for the first two betting rounds, and can either fold, call, bet  $1 \times$  pot, or bet all-in on the last two rounds. The  $\pm$  shows one standard deviation. For Libratus, we list the score against all top humans in aggregate; Libratus beat Dong Kim by 29 with an estimated  $\pm$  of 78.

Table 2 shows ReBeL also converges to an approximate Nash in several versions of Liar’s Dice. Of course, tabular CFR does better than ReBeL when using the same number of CFR iterations, but tabular CFR quickly becomes intractable to run as the game grows in size.

Algorithm	1x4f	1x5f	1x6f	2x3f
Full-game FP	0.012	0.024	0.039	0.057
Full-game CFR	0.001	0.001	0.002	0.002
ReBeL FP	0.041	0.020	0.040	0.020
ReBeL CFR-D	0.017	0.015	0.024	0.017

Table 2: Exploitability of different algorithms of 4 variants of Liar’s Dice: 1 die with 4, 5, or 6 faces and 2 dice with 3 faces. The top two rows represent baseline numbers when a tabular version of the algorithms is run on the entire game for 1,024 iterations. The bottom 2 lines show the performance of ReBeL operating on subgames of depth 2 with 1,024 search iterations. For exploitability computation of the bottom two rows, we averaged the policies of 1,024 playthroughs and thus the numbers are upper bounds on exploitability.

## 9 Conclusions

We present ReBeL, an algorithm that generalizes the paradigm of self-play reinforcement learning and search to imperfect-information games. We prove that ReBeL computes to an approximate Nash equilibrium in two-player zero-sum games and demonstrate that it produces superhuman performance in the benchmark game of heads-up no-limit Texas hold’em.

ReBeL has some limitations that present avenues for future research. Most prominently, the input to its value and policy functions currently grows linearly with the number of infostates in a public state. This is intractable in games such as Recon Chess [46] that have strategic depth but very little common knowledge. ReBeL’s theoretical guarantees are also limited only to two-player zero-sum games.

Nevertheless, ReBeL achieves low exploitability in benchmark games and superhuman performance in heads-up no-limit Texas hold’em while leveraging far less expert knowledge than any prior bot. We view this as a major step toward developing universal techniques for multi-agent interactions.

## Broader Impact

We believe ReBeL is a major step toward general equilibrium-finding algorithms that can be deployed in large-scale multi-agent settings while requiring relatively little domain knowledge. There are numerous potential future applications of this work, including auctions, negotiations, cybersecurity, and autonomous vehicle navigation, all of which are imperfect-information multi-agent interactions.

The most immediate risk posed by this work is its potential for cheating in recreational games such as poker. While AI algorithms already exist that can achieve superhuman performance in poker, these algorithms generally assume that participants have a certain number of chips or use certain bet sizes. Retraining the algorithms to account for arbitrary chip stacks or unanticipated bet sizes requires more computation than is feasible in real time. However, ReBeL can compute a policy for arbitrary stack sizes and arbitrary bet sizes in seconds.

Partly for this reason, we have decided not to release the code for poker. We instead open source our implementation for Liar’s Dice, a recreational game that is not played as competitively by humans. The implementation in Liar’s Dice is also easier to understand and the size of Liar’s Dice can be more easily adjusted, which we believe makes the game more suitable as a domain for research.

## References

- [1] Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. In *Advances in Neural Information Processing Systems*, pages 5360–5370, 2017.
- [2] Robert J Aumann. Agreeing to disagree. *The annals of statistics*, pages 1236–1239, 1976.
- [3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [4] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.
- [5] Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold’em poker is solved. *Science*, 347(6218):145–149, 2015.
- [6] George W Brown. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1):374–376, 1951.
- [7] Noam Brown, Sam Ganzfried, and Tuomas Sandholm. Hierarchical abstraction, distributed equilibrium computation, and post-processing, with application to a champion no-limit texas hold’em agent. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 7–15. International Foundation for Autonomous Agents and Multiagent Systems, 2015.
- [8] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *International Conference on Machine Learning*, pages 793–802, 2019.
- [9] Noam Brown and Tuomas Sandholm. Simultaneous abstraction and equilibrium finding in games. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
- [10] Noam Brown and Tuomas Sandholm. Baby tartanian8: Winning agent from the 2016 annual computer poker competition. In *IJCAI*, pages 4238–4239, 2016.
- [11] Noam Brown and Tuomas Sandholm. Strategy-based warm starting for regret minimization in games. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [12] Noam Brown and Tuomas Sandholm. Safe and nested subgame solving for imperfect-information games. In *Advances in neural information processing systems*, pages 689–699, 2017.
- [13] Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, page eaao1733, 2017.
- [14] Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1829–1836, 2019.

- [15] Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, page eaay2400, 2019.
- [16] Noam Brown, Tuomas Sandholm, and Brandon Amos. Depth-limited solving for imperfect-information games. In *Advances in Neural Information Processing Systems*, pages 7663–7674, 2018.
- [17] Neil Burch, Michael Johanson, and Michael Bowling. Solving imperfect information games using decomposition. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
- [18] Neil Burch, Martin Schmid, Matej Moravcik, Dustin Morill, and Michael Bowling. Aivat: A new variance reduction technique for agent evaluation in imperfect information games. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [19] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Conference on Learning Theory*, pages 6–1, 2012.
- [20] Jilles Steeve Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet. Optimally solving dec-pomdps as continuous-state mdps. *Journal of Artificial Intelligence Research*, 55:443–497, 2016.
- [21] Jakob Foerster, Francis Song, Edward Hughes, Neil Burch, Iain Dunning, Shimon Whiteson, Matthew Botvinick, and Michael Bowling. Bayesian action decoder for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 1942–1951, 2019.
- [22] Sam Ganzfried and Tuomas Sandholm. Potential-aware imperfect-recall abstraction with earth mover’s distance in imperfect-information games. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 682–690, 2014.
- [23] Sam Ganzfried and Tuomas Sandholm. Endgame solving in large imperfect-information games. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 37–45. International Foundation for Autonomous Agents and Multiagent Systems, 2015.
- [24] Sylvain Gelly and David Silver. Combining online and offline knowledge in uct. In *Proceedings of the 24th international conference on Machine learning*, pages 273–280, 2007.
- [25] Andrew Gilpin and Tuomas Sandholm. Optimal rhode island hold’em poker. In *Proceedings of the 20th national conference on Artificial intelligence-Volume 4*, pages 1684–1685, 2005.
- [26] Andrew Gilpin and Tuomas Sandholm. A competitive texas hold’em poker player via automated abstraction and real-time equilibrium computation. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1007. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.
- [27] Eric A Hansen, Daniel S Bernstein, and Shlomo Zilberstein. Dynamic programming for partially observable stochastic games. In *AAAI*, volume 4, pages 709–715, 2004.
- [28] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.
- [29] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- [30] Samid Hoda, Andrew Gilpin, Javier Pena, and Tuomas Sandholm. Smoothing techniques for computing nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2):494–512, 2010.
- [31] Karel Horák and Branislav Bošanský. Solving partially observable stochastic games with public observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2029–2036, 2019.
- [32] Michael Johanson, Nolan Bard, Neil Burch, and Michael Bowling. Finding optimal abstract strategies in extensive-form games. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1371–1379, 2012.
- [33] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

- [34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [35] Vojtěch Kovařík and Viliam Lisý. Problems with the efg formalism: a solution attempt using observations. *arXiv preprint arXiv:1906.06291*, 2019.
- [36] Vojtěch Kovařík, Martin Schmid, Neil Burch, Michael Bowling, and Viliam Lisý. Re-thinking formal models of partially observable multiagent decision making. *arXiv preprint arXiv:1906.11110*, 2019.
- [37] Christian Kroer, Gabriele Farina, and Tuomas Sandholm. Solving large sequential games with the excessive gap technique. In *Advances in Neural Information Processing Systems*, pages 864–874, 2018.
- [38] Christian Kroer, Kevin Waugh, Fatma Kılınç-Karzan, and Tuomas Sandholm. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming*, pages 1–33, 2018.
- [39] Adam Lerer, Hengyuan Hu, Jakob Foerster, and Noam Brown. Improving policies via search in cooperative partially observable games. In *AAAI Conference on Artificial Intelligence*, 2020.
- [40] David S Leslie and Edmund J Collins. Generalised weakened fictitious play. *Games and Economic Behavior*, 56(2):285–298, 2006.
- [41] Viliam Lisý and Michael Bowling. Equilibrium approximation quality of current no-limit poker bots. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [42] 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.
- [43] Matej Moravčík, Martin Schmid, Karel Ha, Milan Hladík, and Stephen J Gaukrodger. Refining subgames in large imperfect information games. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [44] John Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.
- [45] Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644–1658, 2013.
- [46] Andrew J Newman, Casey L Richardson, Sean M Kain, Paul G Stankiewicz, Paul R Guseman, Blake A Schreurs, and Jeffrey A Dunne. Reconnaissance blind multi-chess: an experimentation platform for isr sensor fusion and resource management. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXV*, volume 9842, page 984209. International Society for Optics and Photonics, 2016.
- [47] Frans Adriaan Oliehoek. Sufficient plan-time statistics for decentralized pomdps. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.
- [48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.
- [49] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. 2013.
- [50] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.
- [51] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019.
- [52] Dominik Seitz, Vojtech Kovarik, Viliam Lisý, Jan Rudolf, Shuo Sun, and Karel Ha. Value functions for depth-limited solving in imperfect-information games beyond poker. *arXiv preprint arXiv:1906.06412*, 2019.
- [53] Jack Serrino, Max Kleiman-Weiner, David C Parkes, and Josh Tenenbaum. Finding friend and foe in multi-agent games. In *Advances in Neural Information Processing Systems*, pages 1249–1259, 2019.

- [54] Claude E Shannon. Programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 41(314):256–275, 1950.
- [55] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.
- [56] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.
- [57] Michal Šustr, Vojtěch Kovařík, and Viliam Lisý. Monte carlo continual resolving for online strategy computation in imperfect information games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 224–232. International Foundation for Autonomous Agents and Multiagent Systems, 2019.
- [58] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, pages 2989–2997, 2015.
- [59] Gerald Tesauro. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation*, 6(2):215–219, 1994.
- [60] Ben Van der Genugten. A weakened form of fictitious play in two-person zero-sum games. *International Game Theory Review*, 2(04):307–328, 2000.
- [61] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.
- [62] 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.

## A List of contributions

This paper makes several contributions, which we summarize here.

- **RL+Search in two-player zero-sum imperfect-information games.** Prior work has developed RL+Search for two-player zero-sum perfect-information games. There has also been prior work on learning value functions in fully cooperative imperfect-information games [20] and limited subsets of zero-sum imperfect-information games [31]. However, we are not aware of any prior RL+Search algorithms for two-player zero-sum imperfect-information games in general. We view this as the central contribution of this paper.
- **Alternative to safe search techniques.** Theorem 4 proves that, when doing search at test time with an accurate PBS value function, one can empirically play according to a Nash equilibrium by sampling a random iteration and passing down the beliefs produced by that iteration’s policy. This result applies regardless of how the value function was trained and therefore applies to earlier techniques that use a PBS value function, such as DeepStack [42].
- **Subgame decomposition via CFR-AVE.** We describe the CFR-AVE algorithm in Appendix E. CFR-D [17] is a way to conduct depth-limited solving of a subgame with CFR when given a value function for PBSs. CFR-D is theoretically sound but has certain properties that may reduce performance in a self-play setting. CFR-AVE is a theoretically sound alternative to CFR-D that does not have these weaknesses. However, in order to implement CFR-AVE efficiently, in our experiments we modify the algorithm in a way that is not theoretically sound but empirically performs well in poker. Whether or not this modified form of CFR-AVE is theoretically sound remains an open question.
- **Connection between PBS gradients and infostate values.** Theorem 1 proves that all of the algorithms described in this paper can in theory be conducted using only  $V_1$ , not  $\hat{v}$  (that is, a value function that outputs a single value for a PBS, rather than a vector of values for the infostates in the PBS). While this connection does not have immediate practical consequences, it does point toward a way of deploying the ideas in this paper to settings with billions or more infostates per PBS.
- **Fictitious Linear Optimistic Play (FLOP).** Section D introduces FLOP, a novel variant of Fictitious Play that is inspired by recent work on regret minimization algorithms [14]. We show that FLOP empirically achieves near- $O(\frac{1}{T})$  convergence in the limit in both poker and Liar’s Dice, does far better than any previous variant of FP, and in some domains is a reasonable alternative to CFR.

## B Pseudocode for ReBeL

Algorithm 2 presents ReBeL in more detail.

We define the average of two policies to be the policy that is, in expectation, identical to picking one of the two policies and playing that policy for the entire game. Formally, if  $\pi = \alpha\pi_1 + (1 - \alpha)\pi_2$ , then  $\pi(s_i) = \frac{(x_i^{\pi_1}(s_i)\alpha)\pi_1(s_i) + (x_i^{\pi_2}(s_i)(1-\alpha))\pi_2(s_i)}{x_i^{\pi_1}(s_i)\alpha + x_i^{\pi_2}(s_i)(1-\alpha)}$  where  $x_i^{\pi_1}(s_i)$  is the product of the probabilities for all agent  $i$  actions leading to  $s_i$ . Formally,  $x_i^\pi(s_i)$  of infostate  $s_i = (O_i^0, a_i^0, O_i^1, a_i^1, \dots, O_i^t)$  is  $x_i^\pi(s_i) = \prod_t(a_i^t)$ .

## C Description of Games used for Evaluation

### C.1 Heads-up no-limit Texas hold’em poker (HUNL)

HUNL is the two-player version of no-limit Texas hold’em poker, which is the most popular variant of poker in the world. For each “hand” (game) of poker, each player has some number of chips (the *stack*) in front of them. In our experiments, stack size varies during training between \$5,000 and \$25,000 but during testing is always \$20,000, as is standard in the AI research community. Before play begins, Player 1 commits a *small blind* of \$50 to the pot and Player 2 commits a *big blind* of \$100 to the pot.

---

**Algorithm 2** ReBeL

---

```
function REBEL-LINEAR-CFR-D( $\beta_r, \theta^v, \theta^\pi, D^v, D^\pi$ ) ▷  $\beta_r$  is the current PBS
  while !IS TERMINAL( $\beta_r$ ) do
     $G \leftarrow \text{CONSTRUCTSUBGAME}(\beta_r)$ 
     $\bar{\pi}, \pi^{t_{\text{warm}}} \leftarrow \text{INITIALIZEPOLICY}(G, \theta^\pi)$  ▷  $t_{\text{warm}} = 0$  and  $\pi^0$  is uniform if no warm start
     $G \leftarrow \text{SETLEAFVALUES}(\beta_r, \pi^{t_{\text{warm}}}, \theta^v)$ 
     $v(\beta_r) \leftarrow \text{COMPUTE EV}(G, \pi^{t_{\text{warm}}})$ 
     $t_{\text{sample}} \sim \text{linear}\{t_{\text{warm}} + 1, T\}$  ▷ Probability of sampling iteration  $t$  is proportional to  $t$ 
    for  $t = (t_{\text{warm}} + 1) \dots T$  do
       $\pi^t \leftarrow \text{UPDATEPOLICY}(G, \pi^{t-1})$ 
       $\bar{\pi} \leftarrow \frac{t-1}{t+1} \bar{\pi} + \frac{2}{t+1} \pi^t$ 
       $G \leftarrow \text{SETLEAFVALUES}(\beta_r, \pi^t, \theta^v)$ 
       $v(\beta_r) \leftarrow \frac{t-1}{t+1} v(\beta_r) + \frac{2}{t+1} \text{COMPUTE EV}(G, \pi^t)$ 
      if  $t = t_{\text{sample}}$  then
         $\beta'_r \leftarrow \text{SAMPLELEAF}(G, \pi^t)$  ▷ Sample a leaf PBS according to the new policies
        Add  $\{\beta_r, v(\beta_r)\}$  to  $D^v$  ▷ Add to value net training data
        for  $\beta \in G$  do ▷ Loop over the PBS at every public state in  $G$ 
          Add  $\{\beta, \bar{\pi}(\beta)\}$  to  $D^\pi$  ▷ Add to policy net training data (optional)
         $\beta_r \leftarrow \beta'_r$ 

function SETLEAFVALUES( $\beta, \pi, \theta^v$ )
  if IS LEAF( $\beta$ ) then
    for  $s_i \in \beta$  do ▷ For each infostate  $s_i$  corresponding to  $\beta$ 
       $v(s_i) = \hat{v}(s_i | \beta, \theta^v)$ 
  else
    for  $a \in \mathcal{A}(\beta)$  do
      SETLEAFVALUES( $\mathcal{T}(\beta, \pi, a), \pi, \theta^v$ )

function SAMPLELEAF( $G, \pi$ )
   $i^* \sim \text{unif}\{1, N\}, h \sim \beta_r$  ▷ Sample a history randomly from the root PBS and a random player
  while !IS LEAF( $h$ ) do
     $c \sim \text{unif}[0, 1]$ 
    for  $i = 1 \dots N$  do
      if  $i == i^*$  and  $c < \epsilon$  then ▷ we set  $\epsilon = 0.25$  during training,  $\epsilon = 0$  at test time
        sample an action  $a_i$  uniform random
      else
        sample an action  $a_i$  according to  $\pi_i(s_i(h))$ 
     $h \sim \tau(h, a)$ 
  return  $\beta_h$  ▷ Return the PBS corresponding to leaf node  $h$ 
```

---

Once players commit their blinds, they receive two private cards from a standard 52-card deck. The first of four rounds of betting then occurs. On each round of betting, players take turns deciding whether to fold, call, or raise. If a player folds, the other player receives the money in the pot and the hand immediately ends. If a player calls, that player matches the opponent's number of chips in the pot. If a player raises, that player adds more chips to the pot than the opponent. The initial raise of the round must be at least \$100, and every subsequent raise on the round must be at least as large as the previous raise. A player cannot raise more than either player's stack size. A round ends when both players have acted in the round and the most recent player to act has called. Player 1 acts first on the first round. On every subsequent round, player 2 acts first.

Upon completion of the first round of betting, three *community* cards are publicly revealed. Upon completion of the second betting round, another community card is revealed, and upon completion of the third betting round a final fifth community card is revealed. After the fourth betting round, if no player has folded, then the player with the best five-card poker hand, formed from the player's two private cards and the five community cards, is the winner and takes the money in the pot. In case of a tie, the money is split.

## C.2 Turn endgame hold'em (TEH)

TEH is identical to HUNL except both players automatically call for the first two betting rounds, and there is an initial \$1,000 per player in the pot at the start of the third betting round. We randomize the stack sizes during training to be between \$5,000 and \$50,000 per player. The action space of TEH is reduced to at most three raise sizes ( $0.5 \times \text{pot}$ ,  $1 \times \text{pot}$ , or all-in for the first raise in a round, and  $0.75 \times \text{pot}$  or all-in for subsequent raises), but the raise sizes for non-all-in raises are randomly perturbed by up to  $\pm 0.1 \times \text{pot}$  each game during training. Although we train on randomized stack sizes, bet sizes, and board cards, we measure exploitability on the case of both players having \$20,000, unperturbed bet sizes, and the first four board cards being  $3\heartsuit 7\heartsuit T\spadesuit K\spadesuit$ . In this way we can train on a massive game while still measuring NashConv tractably. Even without the randomized stack and bet sizes, TEH has roughly  $2 \cdot 10^{11}$  infostates.

## C.3 Liar's Dice

Liar's Dice is a two-player zero-sum game in our experiments, though in general it can be played with more than two players. At the beginning of a game each player privately rolls  $d$  dice with  $f$  faces each. After that a betting stage starts where players take turns trying to predict how many dice of a specific kind there are among all the players, e.g., 4 dice with face 5. A player's bid must either be for more dice than the previous player's bid, or the same number of dice but a higher face. The round ends when a player challenges the previous bid (a call of *liar*). If all players together have at least as many dice of the specified face as was predicted by the last bid, then the player who made the bid wins. Otherwise the player who challenged the bid wins. We use the highest face as a *wild* face, i.e., dice with this face count towards a bid for any face.

## D Fictitious Linear Optimistic Play

Fictitious Play (FP) [6] is an extremely simple iterative algorithm that is proven to converge to a Nash equilibrium in two-player zero-sum games. However, in practice it does so at an extremely slow rate. On the first iteration, all agents choose a uniform policy  $\pi_i^0$  and the average policy  $\bar{\pi}_i^0$  is set identically. On each subsequent iteration  $t$ , agents compute a best response to the other agents' average policy  $\pi_i^t = \arg\max_{\pi_i} v_i(\pi_i, \bar{\pi}_{-i}^{t-1})$  and update their average policies to be  $\bar{\pi}_i^t = \frac{t-1}{t} \bar{\pi}_i^{t-1} + \frac{1}{t} \pi_i^t$ . As  $t \rightarrow \infty$ ,  $\bar{\pi}^t$  converges to a Nash equilibrium in two-player zero-sum games.

It has also been proven that a family of algorithms similar to FP known as **generalized weakened fictitious play (GWFP)** also converge to a Nash equilibrium so long as they satisfy certain properties [60, 40], mostly notably that in the limit the policies on each iteration converge to best responses.

In this section we introduce a novel variant of FP we call **Fictitious Linear Optimistic Play (FLOP)** which is a form of GWFP. FLOP is inspired by related variants in CFR, in particular Linear CFR [14]. FLOP converges to a Nash equilibrium much faster than FP while still being an extremely simple algorithm. However, variants of CFR such as Linear CFR and Discounted CFR [14] still converge much faster in most large-scale games.

In FLOP, the initial policy  $\pi_i^0$  is uniform. On each subsequent iteration  $t$ , agents compute a best response to an *optimistic* [19, 49, 58] form of the opponent's average policy in which  $\pi_{-i}^{t-1}$  is given extra weight:  $\pi_i^t = \arg\max_{\pi_i} v_i(\pi_i, \frac{t}{t+2} \bar{\pi}_{-i}^{t-1} + \frac{2}{t+2} \pi_{-i}^{t-1})$ . The average policy is updated to be  $\bar{\pi}_i^t = \frac{t-1}{t+1} \bar{\pi}_i^{t-1} + \frac{2}{t+1} \pi_i^t$ . Theorem 5 proves that FLOP is a form of GWFP and therefore converges to a Nash equilibrium as  $t \rightarrow \infty$ .

**Theorem 5.** *FLOP is a form of Generalized Weakened Fictitious Play.*

*Proof.* Assume that the range of payoffs in the game is  $M$ . Since  $\pi_i^t = \arg\max_{\pi_i} v_i(\pi_i, \frac{t}{t+2} \bar{\pi}_{-i}^{t-1} + \frac{2}{t+2} \pi_{-i}^{t-1})$ , so  $\pi_i^t$  is an  $\epsilon_t$ -best response to  $\bar{\pi}_{-i}^{t-1}$  where  $\epsilon_t < M \frac{2}{t+2}$  and  $\epsilon_t \rightarrow 0$  as  $t \rightarrow \infty$ . Thus, FLOP is a form of GWFP with  $\alpha_t = \frac{2}{t}$ .  $\square$



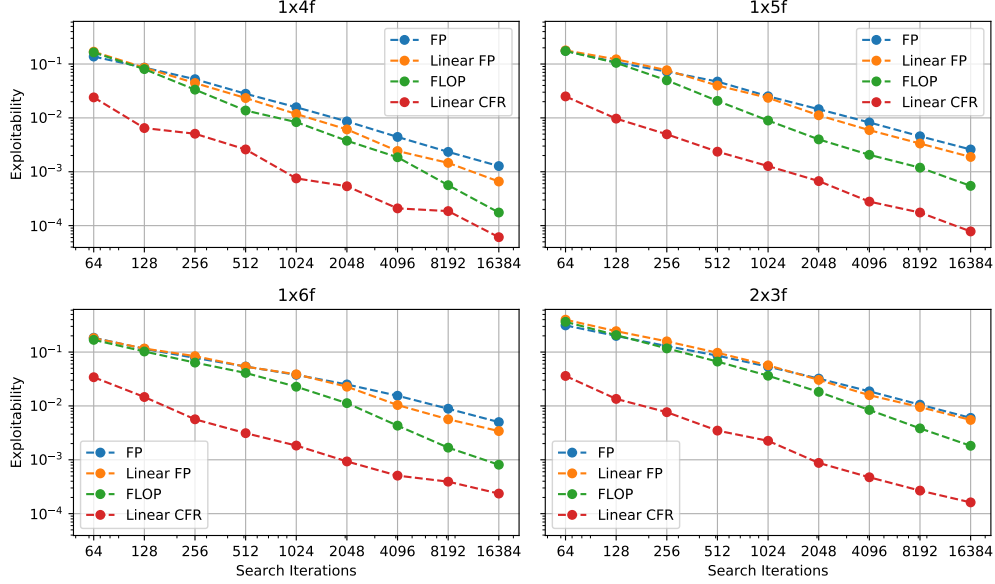


Figure 3: Exploitability of different algorithms of 4 variants of Liar’s Dice: 1 die with 4, 5, or 6 faces and 2 dice with 3 faces. For all games FLOP outperforms Linear FP, but does not match the quality of Linear CFR.

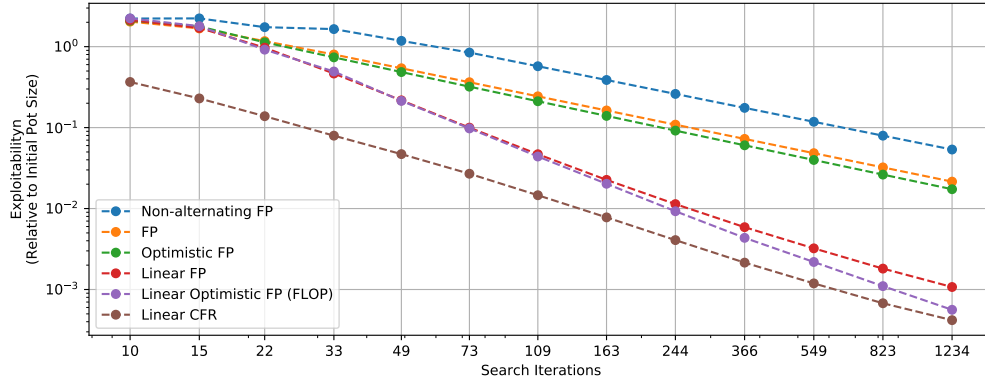


Figure 4: Exploitability of different algorithms for Turn Endgame Hold’em.

## E CFR-AVE: CFR Decomposition using Average Strategy

On each iteration  $t$  of CFR-D, the value of every leaf node  $z$  is set to  $\hat{v}(s_i(z)|\beta_z^{\pi^t})$ . Other than changing the values of leaf nodes every iteration, CFR-D is otherwise identical to CFR. If  $T$  iterations of CFR-D are conducted with a value network that has error at most  $\delta$  for each infostate value, then  $\bar{\pi}^T$  has exploitability of at most  $k_1\delta + k_2/\sqrt{T}$  where  $k_1$  and  $k_2$  are game-specific constants [42].

Since it is the *average* policy profile  $\bar{\pi}^t$ , not  $\pi^t$ , that converges to a Nash equilibrium as  $t \rightarrow \infty$ , and since the leaf PBSs are set based on  $\pi^t$ , the input to the value network  $\hat{v}$  may span the entire domain of inputs even as  $t \rightarrow \infty$ . For example, suppose in a Nash equilibrium  $\pi^*$  the probability distribution at  $\beta_z^{\pi^*}$  was uniform. Then the probability distribution at  $\beta_z^{\pi^t}$  for any individual iteration  $t$  could be *anything*, because regardless of what the probability distribution is, the average over all iterations could still be uniform in the end. Thus,  $\hat{v}$  may need to be accurate over the entire domain of inputs rather than just the subspace near  $\beta_z^{\pi^*}$ .

In **CFR-AVE**, leaf values are instead set according to the *average policy*  $\bar{\pi}^t$  on iteration  $t$ . When a leaf PBS is sampled, the leaf node is sampled with probability determined by  $\pi^t$ , but the PBS itself is defined using  $\bar{\pi}^t$ .

We first describe the tabular form of CFR-D [17]. Consider a game  $G'$  and a depth-limited subgame  $G$ , where both  $G'$  and  $G$  share a root but  $G$  extends only a limited number of actions into the future.

Suppose that  $T$  iterations of a modified form of CFR are conducted in  $G'$ . On each iteration  $t \leq T$ , the policy  $\pi(s_i)$  is set according to CFR for each  $s_i \in G$ . However, for every infostate  $s'_i \in G' \setminus G$ , the policy is set differently than what CFR would call for. At each leaf public state  $s'_{\text{pub}}$  of  $G$ , we solve a subgame rooted at  $\beta_{s'_{\text{pub}}}^{\pi^t}$  by running  $T'$  iterations of CFR. For each  $s'_i$  in the subgame rooted at  $\beta_{s'_{\text{pub}}}^{\pi^t}$ , we set  $\pi^t(s'_i) = \bar{\pi}^T(s'_i)$  (where  $\pi^t(s'_i)$  is the policy for the infostate in  $G$  and  $\bar{\pi}^T(s'_i)$  is the policy for the infostate in the subgame rooted at  $\beta_{s'_{\text{pub}}}^{\pi^t}$ ). It is proven that as  $T' \rightarrow \infty$ , CFR-D converges to a  $O(\frac{1}{\sqrt{T}})$ -Nash equilibrium [17].

CFR-AVE is identical to CFR-D, except the subgames that are solved on each iteration  $t$  are rooted at  $\beta_{s'_{\text{pub}}}^{\bar{\pi}^t}$  rather than  $\beta_{s'_{\text{pub}}}^{\pi^t}$ . Theorem 6 proves that CFR-AVE achieves the same bound on convergence to a Nash equilibrium as CFR-D.

**Theorem 6.** *Suppose that  $T$  iterations of CFR-AVE are run in a depth-limited subgame, where on each iteration  $t \leq T$  the subgame rooted at each leaf PBS  $\beta_{s'_{\text{pub}}}^{\bar{\pi}^t}$  is solved completely. Then  $\bar{\pi}^T$  is a  $\frac{C}{\sqrt{T}}$ -Nash equilibrium for a game-specific constant  $C$ .*

CFR-AVE has a number of potential benefits over CFR-D:

- Since  $\bar{\pi}^t$  converges to a Nash equilibrium as  $t \rightarrow \infty$ , CFR-AVE allows  $\hat{v}$  to focus on being accurate over a more narrow subspace of inputs.
- When combined with a policy network (as introduced in Section 5.3), CFR-AVE may allow  $\hat{v}$  to focus on an even more narrow subspace of inputs.
- Since  $\bar{\pi}^{t+1}$  is much closer to  $\bar{\pi}^t$  than  $\pi^{t+1}$  is to  $\pi^t$ , in practice as  $t$  becomes large one can avoid querying the value network on every iteration and instead recycle the values from a previous iteration. This may be particularly valuable for Monte Carlo versions of CFR.

While CFR-AVE is theoretically sound, we modify its implementation in our experiments to make it more efficient in a way that has not been proven to be theoretically sound. The reason for this is that while the *input* to the value network is  $\beta_{s'_{\text{pub}}}^{\bar{\pi}^t}$  (i.e., the leaf PBS corresponding to  $\bar{\pi}^t$  being played in  $G$ ), the *output* needs to be the value of each infostate  $s_i$  given that  $\pi^t$  is played in  $G$ . Thus, unlike CFR-D and FP, in CFR-AVE there is a mismatch between the input policy and the output policy.

One way to cope with this is to have the input consist of both  $\beta_{s'_{\text{pub}}}^{\bar{\pi}^t}$  and  $\beta_{s'_{\text{pub}}}^{\pi^t}$ . However, we found this performed relatively poorly in preliminary experiments when trained through self play.

Instead, on iteration  $t - 1$  we store the output from  $\hat{v}(s_i | \beta_{s'_{\text{pub}}}^{\bar{\pi}^{t-1}})$  for each  $s_i$  and on iteration  $t$  we set  $v^t(s_i)$  to be  $t\hat{v}(s_i | \beta_{s'_{\text{pub}}}^{\bar{\pi}^t}) - (t - 1)\hat{v}(s_i | \beta_{s'_{\text{pub}}}^{\bar{\pi}^{t-1}})$  (in vanilla CFR). The motivation for this is that  $\pi^t = t\bar{\pi}^t - (t - 1)\bar{\pi}^{t-1}$ . If  $v^t(h) = v^{t-1}(h)$  for each history  $h$  in the leaf PBS, then this modification of CFR-AVE is sound. Since  $v^t(h) = v^{t-1}(h)$  when  $h$  is a full-game terminal node (i.e., it has no actions), this modified form of CFR-AVE is identical to CFR in a non-depth-limited game. However, that is not the case in a depth-limited subgame, and it remains an open question whether this modified form of CFR-AVE is theoretically sound in depth-limited subgames. Empirically, however, we found that it converges to a Nash equilibrium in turn endgame hold'em for every set of parameters (e.g., bet sizes, stack sizes, and initial beliefs) that we tested.

Figure 5 shows the performance of CFR-D, CFR-AVE, our modified form of CFR-AVE, and FP in TEH when using an oracle function for the value network. It also shows the performance of CFR-D, our modified form of CFR-AVE, and FP in TEH when using a value network trained through self-play. Surprisingly, the theoretically sound form of CFR-AVE does worse than CFR-D when using an oracle function. However, the modified form of CFR-AVE does better than CFR-D when using an oracle function and also when trained through self play.

We also trained a model on HUNL with training parameters that were identical to the one reported in Section 8, but using CFR-D rather than CFR-AVE. That model lost to BabyTartanian8 by  $10 \pm 3$  whereas the CFR-AVE model won by  $9 \pm 4$ . The CFR-D model also beat Slumbot by  $39 \pm 6$  whereas the CFR-AVE model won by  $45 \pm 5$ .

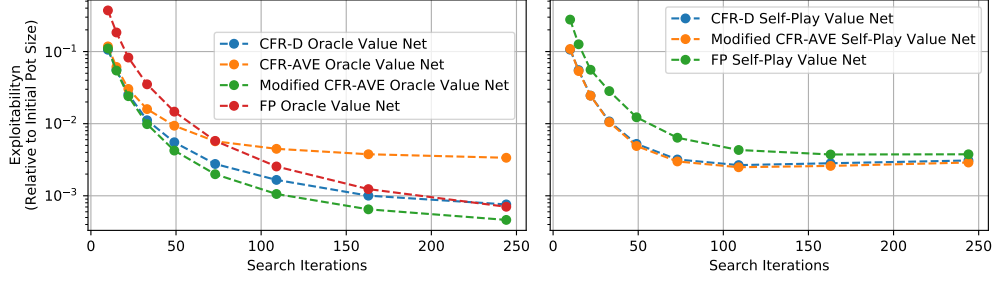


Figure 5: **Left:** comparison of CFR-D, CFR-AVE, modified CFR-AVE, and FP using an oracle value network which returns exact values for leaf PBSs. **Right:** comparison of CFR-D, modified CFR-AVE, and FP using a value network learned through 300 epochs of self play.

### E.1 Proof of Theorem 6

Our proof closely follows from [17] and [42].

*Proof.* Let  $R^t(s_i)$  be the (cumulative) regret of infostates  $s_i$  on iteration  $t$ . We show that the regrets of all infostates in  $G'$  are bounded by  $O(\sqrt{T})$  and therefore the regret of the entire game is bounded by  $O(\sqrt{T})$ .

First, consider the infostates in  $G$ . Since their policies are chosen according to CFR each iteration, their regrets are bounded by  $O(\sqrt{T})$  regardless of the policies played in descendant infostates.

Next consider an infostate  $s_i \in G' \setminus G$ . We prove inductively that  $R^t(s_i) \leq 0$ . Let  $\beta^{\pi^t}$  be the PBS at the root of the subgame containing  $s_i$  in CFR-D, and  $\beta^{\bar{\pi}^t}$  be the PBS at the root of the subgame containing  $s_i$  in CFR-AVE. On the first iteration,  $\beta^{\pi^t} = \beta^{\bar{\pi}^t}$ . Since we assume CFR-AVE computes an exact equilibrium in the subgame rooted at  $\beta^{\bar{\pi}^t} = \beta^{\pi^t}$ , so  $R^t(s_i) = 0$  on the first iteration.

Next, we prove  $R^{t+1}(s_i) \leq R^t(s_i)$ . We define  $a^{*,t}$  as

$$a_i^{*,t} = \operatorname{argmax}_{a_i} \sum_{t'=0}^t v^{t'}(s_i, a_i) \quad (1)$$

By definition of regret,

$$R^{t+1}(s_i) = \sum_{t'=0}^{t+1} (v^{t'}(s_i, a_i^{*,t+1}) - v^{t'}(s_i)) \quad (2)$$

Separating iteration  $t+1$  from the summation we get

$$R^{t+1}(s_i) = \sum_{t'=0}^t (v^{t'}(s_i, a_i^{*,t+1}) - v^{t'}(s_i)) + (v^{t+1}(s_i, a_i^{*,t+1}) - v^{t+1}(s_i)) \quad (3)$$

By definition of  $a_i^{*,t}$  we know  $\sum_{t'=0}^t v^{t'}(s_i, a_i^{*,t+1}) \leq \sum_{t'=0}^t v^{t'}(s_i, a_i^{*,t})$ , so

$$R^{t+1}(s_i) \leq \sum_{t'=0}^t (v^{t'}(s_i, a_i^{*,t}) - v^{t'}(s_i)) + (v^{t+1}(s_i, a_i^{*,t+1}) - v^{t+1}(s_i)) \quad (4)$$

Since  $\sum_{t'=0}^t (v^{t'}(s_i, a_i^{*,t}) - v^{t'}(s_i))$  is the definition of  $R^t(s_i)$  we get

$$R^{t+1}(s_i) \leq R^t(s_i) + (v^{t+1}(s_i, a_i^{*,t+1}) - v^{t+1}(s_i)) \quad (5)$$

Since  $\pi^{t+1} = \pi^{*,t+1}$  in the subgame where  $\pi^{*,t+1}$  is an exact equilibrium of the subgame rooted at  $\beta^{\bar{\pi}^{t+1}}$ , so  $\pi^{t+1}$  is a best response to  $\bar{\pi}^{t+1}$  in the subgame and therefore  $v^{t+1}(s_i, a_i^{*,t+1}) = v^{t+1}(s_i)$ . Thus,

$$R^{t+1}(s_i) \leq R^t(s_i) \quad (6)$$

□

## F Domain Knowledge Leveraged in our Poker AI Agent

The most prominent form of domain knowledge in our ReBeL poker agent is the simplification of the action space during self play so that there are at most 8 actions at each decision point. The bet sizes are hand-chosen based on conventional poker wisdom and are fixed fractions of the pot, though each bet size is perturbed by  $\pm 0.1 \times \text{pot}$  during training to ensure diversity in the training data.

We specifically chose not to leverage domain knowledge that has been widely used in previous poker AI agents:

- All prior top poker agents, including DeepStack [42], Libratus [13], and Pluribus [15], have used *information abstraction* to bucket similar infostates together [32, 22, 7]. Even when computing an exact policy, such as during search or when solving a poker game in its entirety [25, 5], past agents have used *lossless abstraction* in which strategically identical infostates are bucketed together. For example, a flush of spades may be strategically identical to a flush of hearts.

Our agent does not use any information abstraction, whether lossy or lossless. The agent computes a unique policy for each infostate. The agent’s input to its value and policy network is a probability distribution over pairs of cards for each player, as well as all public board cards, the amount of money in the pot relative to the stacks of the players, and a flag for whether a bet has occurred on this betting round yet.

- DeepStack trained its value network on random PBSs. In addition to reducing the dimensionality of its value network input by using information abstraction, DeepStack also sampled PBSs according to a handcrafted algorithm that would sample more realistic PBSs compared to sampling uniform random. We show in Section 8 that training on PBSs sampled uniformly randomly without information abstraction results in extremely poor performance in a value network.

Our agent collects training data purely from self play without any additional heuristics guiding which PBSs are sampled, other than an exploration hyperparameter that was set to  $\epsilon = 0.25$  in all experiments.

- In cases where both players bet all their chips before all board cards are revealed, past poker AIs compute the exact expected value of all possible remaining board card outcomes. This is expensive to do in real time on earlier rounds, so past agents pre-compute this expected value and look it up during training and testing. Using the exact expected value reduce variance and makes learning easier.

Our agent does not use this shortcut. Instead, the agent learns these “all-in” expected values on its own. When both agents have bet all their chips, the game proceeds as normal except neither player is allowed to bet.

- The search space in DeepStack [42] extends to the start of the next betting round, except for the third betting round (out of four) where it instead extends to the end of the game. Searching to the end of the game on the third betting round was made tractable by using information abstraction on the fourth betting round (see above). Similarly, Libratus [12], Modicum [16], and Pluribus [15] all search to the end of the game when on the third betting round. Searching to the end of the game has the major benefit of not requiring the value network to learn values for the end of the third betting round. Thus, instead of the game being three “levels” deep, it is only two levels deep. This reduces the potential for propagation of errors.

Our agent always solves to the end of the current betting round, regardless of which round it is on.

- The depth-limited subgames in DeepStack extended to the start of the next betting round on the second betting round. On the first betting round, it extended to the end of the first betting round for most of training and to the start of the next betting round for the last several CFR iterations. Searching to the start of the next betting round was only tractable due to the abstractions mentioned previously and due to careful optimizations, such as implementing CFR on a GPU.

Our agent always solves to the end of the current betting round regardless of which round it is on. We implement CFR only on a single-thread CPU and avoid any abstractions. Since a subgame starts at the beginning of a betting round and ends at the start of the next betting

round, our agent must learn six “layers” of values (end of first round, start of second round, end of second round, start of third round, end of third round, start of fourth round) compared to three for DeepStack (end of first round, start of second round, start of third round).

- DeepStack used a separate value network for each of the three “layers” of values (end of first round, start of second round, start of third round). Our agent uses a single value network for all situations.

## G Hyper parameters

In this section we provide details of the value and policy networks and the training procedures.

We approximate the value and policy functions using artificial neural networks. The input to the value network consists of three components for both games: agent index, representation of the public state, and a probability distribution over infostates for both agents. For poker, the public state representation consists of the board cards and the common pot size divided by stack size; for Liar’s Dice it is the last bid and the acting agent. The output of the network is a vector of values for each possible infostate of the indexed agent, e.g., each possible poker hand she can hold.

We trained a policy network only for poker. The policy network state representation additionally contains pot size fractions for both agents separately as well as a flag for whether there have been any bets so far in the round. The output is a probability distribution over the legal actions for each infostate.

As explained in section 7 we use Multilayer perceptron with GeLU [29] activation functions and LayerNorm [3] for both value and policy networks.

For poker we represent the public state as a concatenation of a vector of indices of the board cards, current pot size relative to the stack sizes, and binary flag for the acting player. The size of the full input is

$$1(\text{agent index}) + 1(\text{acting agent}) + 1(\text{pot}) + 5(\text{board}) + 2 \times 1326(\text{infostate beliefs})$$

We use card embedding for the board cards similar to [8] and then apply MLP. Both the value and the policy networks contain 6 hidden layers with 1536 layers each. For all experiments we set the probability to explore a random action to  $\epsilon = 25\%$  (see Section 5.2). To store the training data we use a simple circular buffer of size 12M and sample uniformly. Since our action abstraction contains at most 9 legal actions, the size of the target vector for the policy network is 9 times bigger than one used for the value network. In order to make it manageable, we apply linear quantization to the policy values. As initial data is produced with a random value network, we remove half of the data from the replay buffer after 20 epochs.

For the full game we train the network with Adam optimizer with learning rate  $3 \times 10^{-4}$  and halved the learning rate every 800 epochs. One epoch is 2,560,000 examples and the batch size 1024. We used 90 DGX-1 machines, each with  $8 \times 32\text{GB}$  Nvidia V100 GPUs for data generation. We report results after 1,750 epochs. For TEH experiments we use higher initial learning rate  $4 \times 10^{-4}$ , but halve it every 100 epochs. We report results after 300 epochs.

For Liar’s Dice we represent the state as a concatenation of a one hot vector for the last bid and binary flag for the acting player. The size of the full input is

$$1(\text{agent index}) + 1(\text{acting agent}) + n_{\text{dice}}n_{\text{faces}}(\text{last bid}) + 2n_{\text{faces}}^{n_{\text{dice}}}(\text{infostate beliefs}).$$

The value network contains 2 hidden layers with 256 layers each. We train the network with Adam optimizer with learning rate  $3 \times 10^{-4}$  and halved the learning rate every 400 epochs. One epoch is 25,600 examples and the batch size 512. During both training and evaluation we run the search algorithm for 1024 iterations. We use single GPU for training and 60 CPU threads for data generation. We trained the network for 1000 epochs. To reduce the variance in RL+Search results, we evaluated the three last checkpoints and reported averages in table 2.

### G.1 Human Experiments for HUNL

We evaluated our HUNL agent against Dong Kim, a top human professional specializing in HUNL. Kim was one of four humans that played against Libratus [13] in the man-machine competition which

Libratus won. Kim lost the least to Libratus. However, due to high variance, it is impossible to statistically compare the performance of the individual humans that participated in the competition.

A total of 7,500 hands were played between Kim and the bot. Kim was able to play from home at his own pace on any schedule he wanted. He was also able to play up to four games simultaneously against the bot. To incentivize strong play, Kim was offered a base compensation of  $\$1 \pm \$0.05x$  for each hand played, where  $x$  signifies his average win/loss rate in terms of big blinds per hundred hands played. Kim was guaranteed a minimum of \$0.75 per hand and could earn no more than \$2 per hand. Since final compensation was based on the variance-reduced score rather than the raw score, Kim was not aware of his precise performance during the experiment.

The bot played at an extremely fast pace. No decision required more than 5 seconds, and the bot on average plays faster than 2 seconds per hand in self play. To speed up play even further, the bot cached subgames it encountered on the preflop, flop, and turn. When the same subgame was encountered again, it would simply reuse the solution it had already computed previously.

Kim's variance-reduced score, which we report in Section 8, was a loss of  $165 \pm 69$  where the  $\pm$  indicates one standard error. His raw score was a loss of  $358 \pm 188$ .

## H Proofs Related to Value Functions (Theorems 1 and 2)

We start by proving some preliminary Lemmas. For simplicity, we will sometimes prove results for only one player, but the results hold WLOG for both players.

For some policy profile  $\pi = (\pi_1, \pi_2)$ , let  $v_i^\pi(s_1|\beta) : \mathcal{B} \rightarrow \mathbb{R}^{|\mathcal{S}_i|}$  be a function that takes as input a PBS and outputs infoset values for player  $i$  at infoset  $s_1$ .

**Lemma 1.** *For fixed  $\beta$  and  $\pi_2$ ,  $v_1^{(\pi_1, \pi_2)}(s_1|\beta)$  is identical for any  $\pi_1$  that is a BR to  $\pi_2$  if  $\beta_1(s_1) > 0$ .*

*Proof.*  $\pi_1^*$  is a BR therefore it must maximize  $V_1 = \sum_{s_1} p(s_1) v_1^\pi(s_1)$ . It can only do so by achieving the unique maximum at each infoset  $s_1$  that occurs with positive probability.  $\square$

**Lemma 2.** *Let  $V_1^{\pi_2}(\beta)$  be player 1's BR value at  $\beta$  assuming that player 2 plays  $\pi_2$ .  $V_1^{\pi_2}(\beta)$  is linear in  $\beta_1$ .*

*Proof.* This follows directly from Lemma 1 along with the definition of  $V_1$ ,

$$V_1^{\pi_2}(\beta) = \sum_{s_1 \in \mathcal{S}_1(s_{\text{pub}})} \beta_1(s_1) v_1(s_1|\beta, (BR(\pi_2), \pi_2))$$

.

$\square$

**Lemma 3.**  *$V_1(\beta) = \min_{\pi_2} V_1^{\pi_2}(\beta)$ , and the set of  $\pi_2$  that attain  $V_1(\beta)$  at  $\beta_0$  are precisely the Nash equilibrium policies at  $\beta_0$ . This also implies that  $V_1(\beta)$  is concave.*

*Proof.* By definition, the Nash equilibrium at  $\beta$  is the minimum among all choices of  $\pi_2$  of the value to player 1 of her BR to  $\pi_2$ . Any  $\pi_2$  that achieves this NE value when playing a BR is a NE policy.

From Lemma 2, we know that each  $V_1^{\pi_2}(\beta)$  is linear, which implies that  $V_1(\beta)$  is concave since any function that is the minimum of linear functions is concave.  $\square$

**Lemma 4.** *At any  $\beta$ , the set of maps  $v_1 : \mathcal{S}_1 \rightarrow \mathbb{R}$  corresponding to Nash equilibrium policies  $\pi^*$  forms a convex set.*

*Proof.* A mixture of Nash equilibrium policy profile is a coarse correlated equilibrium, which means it's a Nash equilibrium since the game is two-player zero-sum. Therefore the set of Nash equilibrium policies is convex on the simplex.

Now, consider the map from infosets to values, using a normal form representation of the subgame:

$$v_1^{\pi^*}(s_1|\beta) = \sum_{h \in s_1} p(h|\beta) \pi_1^*(a_1) \pi_2^*(a_2) v_1(h|a_1, a_2) \quad (7)$$

This map is continuous in  $\pi^*$ , so the set of maps must also be convex.  $\square$

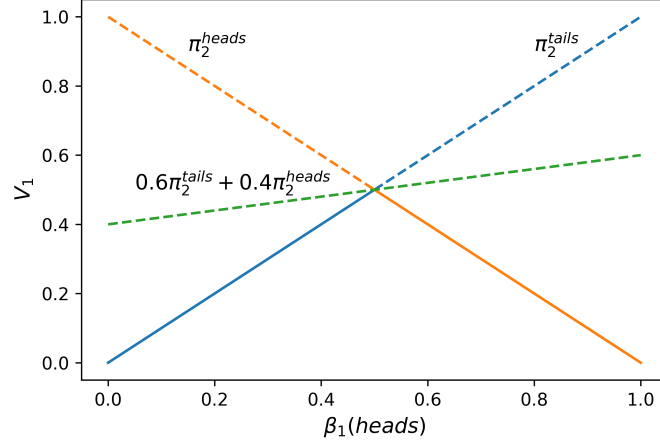


Figure 6: Illustration of Lemma 3. In this simple example, the subgame begins with some probability  $\beta(\text{heads})$  of a coin being heads-up, which player 1 observes. Player 2 then guesses if the coin is heads or tails, and wins if he guesses correctly. The payoffs for Player 2's pure strategies are shown as the lines marked  $\pi_2^{\text{heads}}$  and  $\pi_2^{\text{tails}}$ . The payoffs for a mixed strategy is a linear combination of the pure strategies. The value for player 1 is the minimum among all the lines corresponding to player 2 strategies, denoted by the solid lines.

Now we can turn to proving the Theorem.

Consider a function  $\tilde{V}_1$  that is an extension of  $V_1$  to unnormalized probability distributions over  $S_1$  and  $S_2$ ; i.e.  $\tilde{V}_i((s_{\text{pub}}, b_1, b_2)) = V_i((s_{\text{pub}}, b_1/|b_1|_1, b_2/|b_2|_1))$ .  $\tilde{V}_i = V_i$  on the simplex of valid beliefs, but we extend it in this way to  $\mathbb{R}_{\geq 0}^{|S_1|} \setminus \vec{0}$  so that we can consider gradients w.r.t.  $p(s_1)$ .

We will use the term ‘supergradient’ to be the equivalent of the subgradient for concave functions. Formally,  $g$  is a supergradient of concave function  $F$  at  $x_0$  iff for any  $x$  in the domain of  $F$ ,

$$F(x) - F(x_0) \leq g \cdot (x - x_0).$$

Also,  $\text{superg}(F) = -\text{subg}(-F)$ .

**Theorem** (Restatement of Theorem 1). *For any belief  $\beta_1$  and any supergradient  $\bar{g}$  of  $\tilde{V}_1(\beta)$  with respect to  $\beta_1$ ,*

$$v_1^{\pi^*}(s_1|\beta) = V_1(\beta) + \bar{g} \cdot \hat{s}_1 \quad (8)$$

for some Nash equilibrium policy  $\pi^*$ , where  $\hat{s}_1$  is the unit vector in direction  $s_1$ .

*Proof.* Lemma 3 shows that  $V_1(\beta)$  is a concave function of  $\beta$ , and its extension  $\tilde{V}$  off the simplex is constant perpendicular to the simplex, so  $\tilde{V}$  is concave as well. Therefore the notion of a supergradient is well-defined.

Furthermore,  $V_1$  is the minimum of a set of linear functions  $V_1^{\pi_2}$  (Lemma 3), so at each point  $\beta$ ,  $V_1$  is equal to  $V_1^{\pi_2}$  for one or more policies  $\pi_2$  which are exactly the set of equilibrium policies at  $\beta$ . The gradient of  $V_1^{\pi_2}$  is

$$\nabla_{\beta_1} V_1^{\pi_2}(\beta) = \nabla_{\beta_1} \sum_{s_1 \in S_1(s_{\text{pub}})} \beta_1(s_1) v_1^{(BR(\pi_2), \pi_2)}(s_1|\beta) \quad (9)$$

$$= \sum_{s_1 \in S_1(s_{\text{pub}})} \hat{s}_1 v_1^{(BR(\pi_2), \pi_2)}(s_1|\beta) \quad (10)$$

$$(11)$$

If there is only a single  $V_1^{\pi^2}(\beta)$  plane that intersects  $V_1(\beta)$ , then  $V_1$  lies on this plane and has a single supergradient which is simply the gradient of  $V_1^{\pi^2}$  at this point<sup>9</sup>.

Otherwise,  $V_1(\beta)$  lies at an ‘edge’ defined by the intersection of planes corresponding to  $V_1^{\pi^2}(\beta)$  for different equilibrium policies. The tangent plane for any supergradient at  $\beta_1$  lies within the convex hull of these intersecting planes. By Lemma 4, the set of  $V_1^{\pi^2}$  planes is convex so any plane in this convex hull corresponds to the value for some equilibrium  $\pi^*$ . Therefore, any supergradient of  $V$  corresponds to a  $\nabla_{\beta_1} V_1^{\pi^*}(\beta_1)$  for some NE  $\pi^*$  in the subgame.

Finally, let’s compute  $g \cdot \hat{s}_1$  at some  $\beta_1$  on the simplex (i.e.  $|\beta_1|_1 = 1$ ).

$$g = \nabla_{\beta_1/|\beta_1|_1} V_1^{\pi^*}(s_{\text{pub}}, \beta_1/|\beta_1|, \beta_2) \cdot \frac{d}{d\beta_1} \left( \frac{\beta_1}{|\beta_1|_1} \right) \quad (\text{chain rule}) \quad (12)$$

$$= \left( \sum_{s'_1 \in S_1(s_{\text{pub}})} \hat{s}'_1 v_1^{\pi^*}(s'_1|\beta) \right) \cdot (|\beta_1|_1 - \beta_1)/(|\beta_1|_1)^2 \quad (\text{Eq. 11}) \quad (13)$$

$$= \left( \sum_{s'_1 \in S_1(s_{\text{pub}})} \hat{s}'_1 v_1^{\pi^*}(s'_1|\beta) \right) \cdot (1 - \beta_1) \quad (\text{since } |\beta_1|_1 = 1) \quad (14)$$

$$= \sum_{s'_1 \in S_1(s_{\text{pub}})} \hat{s}'_1 v_1^{\pi^*}(s'_1|\beta) - \sum_{s'_1 \in S_1(s_{\text{pub}})} \beta_1(s'_1) v_1^{\pi^*}(s'_1|\beta) \quad (15)$$

$$= \sum_{s'_1 \in S_1(s_{\text{pub}})} \hat{s}'_1 v_1^{\pi^*}(s'_1|\beta) - V_1(\beta) \quad (16)$$

$$g \cdot \hat{s}_1 = v_1^{\pi^*}(s_1|\beta) - V_1(\beta) \quad (17)$$

And we’re done. □

**Theorem** (Restatement of Theorem 2). *Let  $X$  be the vector of infostate EVs in PBS  $\beta$  corresponding to minimax policy profile  $\pi_X^*$ , and let  $Y$  be the vector of infostate EVs in  $\beta$  corresponding to minimax policy profile  $\pi_Y^*$ . Then  $\lambda X + (1 - \lambda)Y$  is the vector of infostate EVs in  $\beta$  corresponding to minimax policy profile  $\lambda\pi_X^* + (1 - \lambda)\pi_Y^*$  for  $0 \leq \lambda \leq 1$ .*

*Proof.* We focus on a single infostate EV,  $v_1(s_1|\beta, \pi^*)$ , and consider a normal form representation of the subgame.

$$v_1^{\lambda\pi_X^* + (1-\lambda)\pi_Y^*}(s_1|\beta) = \sum_{h \in s_1} p(h|\beta) (\lambda\pi_X^*(a) + (1 - \lambda)\pi_Y^*(a)) v_1(h|a) \quad (18)$$

$$= \lambda \sum_{h \in s_1} p(h|\beta) \pi_X^*(a) v_1(h|a) + (1 - \lambda) \sum_{h \in s_1} p(h|\beta) \pi_Y^*(a) v_1(h|a) \quad (19)$$

$$= \lambda v_1^{\pi_X^*}(s_1|\beta) + (1 - \lambda) v_1^{\pi_Y^*}(s_1|\beta) \quad (20)$$

This mixed joint policy can be played independently by each agent in a two-player zero-sum game, since all coarse correlated equilibria are Nash equilibria. The interpolated policy is a minimax (Nash) strategy, due to Lemma 4. □

## I Proofs Related to Subgame Solving (Theorems 3 and 4)

**Theorem** (Restatement of Theorem 3). *Consider an idealized value approximator that returns the most recent sample of the value for sampled PBSs, and 0 otherwise. Running Algorithm 1*

<sup>9</sup>The supergradient of a differentiable function is only equal to its gradient in the interior, which is why we exclude the boundary from the result, i.e. we only consider  $\mathbb{R}_{>0}^{|S_1|}$ . This could be corrected with a more detailed proof, but we don’t care about the boundary since CFR-AVE never assigns a probability of exactly 0 to any state.



with  $T$  iterations of CFR in each subgame will, after a finite amount of time, produce a value approximator that produces values that correspond to a  $\frac{C}{\sqrt{T}}$ -equilibrium policy for any PBS that could be encountered during play, where  $C$  is a game-dependent constant.

*Proof.* CFR [62] is an iterated self play algorithm whose average policy across iterations converges to a Nash equilibrium. The key idea behind CFR is that it decomposes the regret minimization in the full game into independent regret minimization problems at each information state. At each infostate  $I$ , CFR minimizes the regret over the *counterfactual value*, that is the EV of taking action  $a$  at  $I$  weighted by the probability of reaching  $I_i$  assuming player  $i$  plays to reach  $I$  and the opponent and chance play their policies at iteration  $t$ . The central result of [62] is that the total regret  $R_i^T$  in the game is bounded by the sum of the counterfactual regrets at each infostate  $R_i^T(I)$ . [62] then proposes an independent regret matching policy [28] of

$$\pi_i^{t+1}(s_i, a_i) = \begin{cases} \frac{\max\{0, R_i^t(s_i, a_i)\}}{\sum_{a'_i \in \mathcal{A}_i(s_i)} \max\{0, R_i^t(s_i, a'_i)\}} & \text{if } \sum_{a'_i \in \mathcal{A}_i(s_i)} \max\{0, R_i^t(s_i, a'_i)\} > 0 \\ \frac{1}{|\mathcal{A}_i(s_i)|} & \text{otherwise} \end{cases} \quad (21)$$

at each infostate, whose external regret after  $T$  iterations is bounded by  $O(1/\sqrt{T})$ . This leads to the CFR bound

$$R_i^T \leq \Delta |\mathcal{I}_i| \sqrt{|\mathcal{A}_i|} / \sqrt{T},$$

where  $\Delta$  is the range of payoffs,  $|\mathcal{I}_i|$  is the number of infostates, and  $|\mathcal{A}_i|$  is the max number of actions for player  $i$ .

A crucial property of CFR is that each regret minimization at  $I$  only depends on the counterfactual values of each action at  $I$ . It doesn't matter exactly what policy is performed at other infostates as long as they have low regret.

Suppose we compute an  $\epsilon$ -Nash equilibrium in a PBS  $\beta^{\pi^*}$ . Then the values for the PBS correspond to an (average) policy in  $\beta^{\pi^*}$  that achieves at most  $\epsilon$  regret in  $I$  at time  $t$ .

Consider CFR run in a depth-limited subgame  $\mathcal{G}$ . Let  $I_i^L$  be the infostates in leaf  $L$ , and  $I_i^{\mathcal{G}*}$  be the infostates of  $\mathcal{G}$  not in any leaf subgame. If the total regret at each leaf PBS  $\beta_L^{\pi^*}$  at each iteration  $t$  is bounded by  $|I_i^L|/\sqrt{T}$  then the total regret in  $\mathcal{G}$  will be bounded by

$$R_{i,\mathcal{G}}^T \leq |I_i^{\mathcal{G}*}|/\sqrt{T} + \frac{1}{T} \sum_{L \in \mathcal{G}} \sum_{t=1}^T R_{i,\beta_L^{\pi^*}}^T \quad (22)$$

$$= |I_i^{\mathcal{G}*}|/\sqrt{T} + \frac{1}{T} \sum_{L \in \mathcal{G}} \sum_{t=1}^T |I_i^L|/\sqrt{T} \quad (23)$$

$$= |I_i^{\mathcal{G}*}|/\sqrt{T} + \sum_{L \in \mathcal{G}} |I_i^L|/\sqrt{T} \quad (24)$$

$$= |I_i^{\mathcal{G}}|/\sqrt{T}. \quad (25)$$

In other words, if the  $O(1/\sqrt{T})$  regret bound holds at each leaf PBS  $\beta_L^{\pi^*}$  encountered during the search in  $\beta^{\mathcal{G}}$ , then the regret bound also holds for the values computed in  $\beta$ . So now we must show inductively that valid bounds are computed for each relevant PBS that may be encountered during play.

Consider the naive algorithm that at each leaf node in a depth-limited subgame, recursively runs the same CFR procedure in each leaf subgame  $\beta_L^t$ . This algorithm would clearly obey the regret bound in Equation 25 by sequentially solving  $O(t^{|\mathcal{S}_{pub}|})$  PBSs. But what about the sampling approach in Algorithm 1?

Every PBS solved during this naive scheme can be specified by a pair of a public state  $s_{pub}$  and a sequence  $\tau = (t_1, t_2, \dots, t_k)$ , denoting that this is the PBS at  $s_{pub}$  with beliefs that arise from  $\pi^{t_1}$  at the root PBS,  $\pi^{t_2}$  at the second subgame on the path to  $s_{pub}$ , and so on. We call  $\tau$  an **iter-sequence**.

We define the "natural ordering" of iter-sequences to be the one that places all suffixes of  $\tau$  before  $\tau$ , and orders all prefixes lexicographically. E.g. for  $T = 2$  and depth of 3, the natural ordering would be  $(1, 1, 1), (1, 1, 2), (1, 1), (1, 2, 1), (1, 2, 2), (1, 2), (1), (2, 1, 1), (2, 1, 2), (2, 1), (2, 2, 1), (2, 2, 2), (2, 2), (2), (2)$ .

Consider a PBS  $\beta$  with some iter-sequence  $\tau$ . Suppose the value function for every PBS with an iter sequence  $\tau' < \tau$  is "valid" at iteration  $j$  of Algorithm 1. Then CFR will be correct in all subgames leading to  $\beta$  up to the relevant iteration. So with positive probability, at iteration  $j$  of Algorithm 1, the sequence of leaf PBSs leading to  $\beta$  will be reached, and CFR will be evaluated in  $\beta$ . Furthermore, the PBSs for all subgames of  $\beta$  are suffixes in the natural ordering, so values for all leaf nodes in  $\beta$  will be "valid". Therefore, with positive probability, a correct value of  $\beta$  will be inserted into the value function at this iteration of Algorithm 1. And all future computations of the value of  $\beta$  will also be correct, so the value of  $\beta$  will always be valid after this point.

Order all PBSs encountered during this procedure  $(\beta_1, \dots, \beta_N)$  by natural ordering of their iter-sequences. Suppose that on iter  $j$  of Algorithm 1, the value function is valid for  $(\beta_1, \dots, \beta_k)$ . Then  $P(\beta_{k+1} \text{ sampled on iter } m)$  is positive and independent on each iter  $m > j$ . Therefore by the second Borel-Cantelli Lemma,  $\lim_{M \rightarrow \infty} Pr(\beta_{k+1} \text{ sampled on some iter } j < m < M) = 1$ . So each  $\beta$  will eventually be sampled and produce a valid value estimate.

The only thing left to show is that any  $\beta$  encountered during play against the final policies generated by this procedure will be in the set of  $\{\beta\}$  computed by Algorithm 1. The set of possible test-time PBSs consist of those where the agent plays  $\pi^T$  and the opponent plays an arbitrary policy. As described in Section 5.2, leaf PBSs are sampled at a random  $t \leq T$  for a public state reached by one player playing  $\pi^t$  and the other playing a uniform policy with probability  $\epsilon$  (and  $\pi^t$  otherwise). So it will sample every  $\beta_L^{\pi^T}$  for any  $L$  as long as it's in the support of  $\pi^T$  for at least one player. This is a superset of all leaf nodes that may be encountered when the agent plays  $\pi^T$  at test time.  $\square$

**Theorem** (Restatement of Theorem 4). *If Algorithm 1 is run at test time with no off-policy exploration, a value network that has error at most  $\delta$  for any leaf PBS, and with  $T$  iterations of CFR being used to solve subgames, then the algorithm plays a  $(\delta C_1 + \frac{\delta C_2}{\sqrt{T}})$ -Nash equilibrium, where  $C_1, C_2$  are game-specific constants.*

*Proof.* We prove the theorem inductively. Consider first a subgame near the end of the game that is not depth-limited. I.e., it has no leaf nodes. Clearly, the policy  $\pi^*$  that Algorithm 1 using CFR plays in expectation is a  $\frac{k_1}{\sqrt{T}}$ -Nash equilibrium for game-specific constant  $k_1$  in this subgame.

Rather than play the average policy over all  $T$  iterations  $\bar{\pi}^T$ , one can equivalently pick a random iteration  $t \sim \text{uniform}\{1, T\}$  and play according to  $\pi^t$ , the policy on iteration  $t$ . This algorithm is also a  $\frac{k_1}{\sqrt{T}}$ -Nash equilibrium in expectation.

Next, consider a depth-limited subgame  $G$  such that for any leaf PBS  $\beta^t$  on any CFR iteration  $t$ , the policy that Algorithm 1 plays in the subgame rooted at  $\beta^t$  is in expectation a  $\delta$ -Nash equilibrium in the subgame. If one computes a policy for  $G$  using tabular CFR-D [17] (or, as discussed in Section E, using CFR-AVE), then by Theorem 2 in [17], the average policy over all iterations is  $k_2\delta + \frac{k_3}{\sqrt{T}}$ -Nash equilibrium.

Just as before, rather than play according to this average policy  $\bar{\pi}^T$ , one can equivalently pick a random iteration  $t \sim \text{uniform}\{1, T\}$  and play according to  $\pi^t$ . Doing so would also result in a  $k_2\delta + \frac{k_3}{\sqrt{T}}$ -Nash equilibrium in expectation. This is exactly what Algorithm 1 does.

Since there are a finite number of "levels" in a game, which is a game-specific constant, Algorithm 1 plays according to a  $\delta C_1 + \frac{\delta C_2}{\sqrt{T}}$ -Nash equilibrium.  $\square$

## J CFR Warm Start Algorithm Used

Our warm start technique for CFR is based on [11], which requires only a policy profile to warm start CFR soundly. That technique computes a "soft" best response to the policy profile, which results in

instantaneous regrets for each infostate. Those instantaneous regrets are scaled up to be equivalent to some number of CFR iterations. However, that technique requires careful parameter tuning to achieve good performance in practice.

We instead use a simplified warm start technique in which an exact best response to the policy profile is computed. That best response results in instantaneous regrets at each infostate. Those regrets are scaled up by a factor of 15 to imitate 15 CFR iterations. Similarly, the average policy effectively assumes that the warm start policy was played for the first 15 iterations of CFR. CFR then proceeds as if 15 iterations have already occurred.