Reduced Space and Faster Convergence in Imperfect-Information Games via Pruning

Noam Brown [1] **Tuomas Sandholm** [1]

Abstract

Iterative algorithms such as *Counterfactual Regret Minimization* (CFR) are the most popular way to solve large zero-sum imperfect-information games. In this paper we introduce *Best-Response Pruning* (BRP), an improvement to iterative algorithms such as CFR that allows poorly-performing actions to be temporarily pruned. We prove that when using CFR in zero-sum games, adding BRP will asymptotically prune any action that is not part of a best response to some Nash equilibrium. This leads to provably faster convergence and lower space requirements. Experiments show that BRP results in a factor of 7 reduction in space, and the reduction factor increases with game size.

1. Introduction

Imperfect-information extensive-form games model strategic multi-step scenarios between agents with hidden information, such as auctions, security interactions (both physical and virtual), negotiations, and military situations. Typically in imperfect-information games, one wishes to find a Nash equilibrium, which is a profile of strategies in which no player can improve her outcome by unilaterally changing her strategy. A linear program can find an exact Nash equilibrium in two-player zero-sum games containing fewer than about 10^8 nodes (Gilpin & Sandholm, 2007). For larger games, iterative algorithms are used to converge to a Nash equilibrium. There are a number of such iterative algorithms (Heinrich et al., 2015; Nesterov, 2005; Hoda et al., 2010; Pays, 2014; Kroer et al., 2015), the most popular of which is *Counterfactual Regret Minimization* (CFR) (Zinkevich et al., 2007). CFR minimizes regret independently at each decision point in the game. CFR+, a variant of CFR, was used to essentially solve Limit Texas Hold'em, the largest imperfect-information game ever to be essentially solved (Bowling et al., 2015).

Both computation time and storage space are difficult challenges when solving large imperfect-information games. For example, solving Limit Texas Hold'em required nearly 8 million core hours and a complex, domain-specific streaming compression algorithm to store the 262 TiB of uncompressed data in only 10.9 TiB. This data had to be repeatedly decompressed from disk into memory and then compressed back to disk in order to run CFR+ (Tammelin et al., 2015).

In certain situations, pruning can be applied to speed up the traversal of the game tree in iterative algorithms (Lanctot et al., 2009; Brown & Sandholm, 2015a; Brown et al., 2017). However, these past pruning techniques do not reduce the space needed to solve a game and lack theoretical guarantees for improved performance.

In this paper we introduce *Best-Response Pruning* (BRP)[1], a new form of pruning for iterative algorithms such as CFR in large imperfect-information games. BRP leverages the fact that in iterative algorithms we are typically interested in performance against the opponent's average strategy over all iterations, and that the opponent's average strategy cannot change faster than a rate of $\frac{1}{t}$, where t is the number of iterations conducted so far. Thus, if part-way through a run one of our actions has done very poorly relative to other available actions against the opponent's average strategy, then after just a few more iterations the opponent's average strategy cannot change sufficiently for the poorly-performing action to now be doing well against the opponent's updated average strategy. In fact, we can bound how much an action's performance can improve over any number of iterations against the opponent's average strategy. So long as the upper bound on that performance is still not competitive with the other actions, then we can safely ignore the poorly-performing action.

BRP provably reduces the computation time needed to solve imperfect-information games. Additionally, a primary advantage of BRP is that in addition to faster convergence,

[1]Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA. Correspondence to: Noam Brown <noamb@cs.cmu.edu>, Tuomas Sandholm <sandholm@cs.cmu.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[1]Earlier versions of this paper referred to Best-Response Pruning as *Total Regret-Based Pruning* (Total RBP)

it also reduces the *space* needed over time. Specifically, once pruning begins on a branch, BRP discards the memory allocated on that branch and does not reallocate the memory until pruning ends and the branch cannot immediately be pruned again. In Section 3.1, we prove that after enough iterations of CFR are completed, space for certain pruned branches will *never* need to be allocated again. Specifically, we prove that when using BRP it is asymptotically only necessary to store data for parts of the game that are reached with positive probability in a best response to a Nash equilibrium. This is extremely advantageous when solving large imperfect-information games, which are often constrained by space and in which the set of best response actions may be orders of magnitude smaller than the size of the game (Schmid et al., 2014).

While BRP still requires enough memory to store the entire game in the early iterations, recent work has shown that these early iterations can be skipped in CFR, and possibly other iterative algorithms, by first solving a low-memory abstraction of the game and then using its solution to warm start CFR in the full game (Brown & Sandholm, 2016). BRP's reduction in space is also helpful to the *Simultaneous Abstraction and Equilibrium Finding* (SAEF) algorithm (Brown & Sandholm, 2015b), which starts CFR with a small abstraction of the game and progressively expands the abstraction while also solving the game. SAEF's space requirements increase the longer the algorithm runs, and may eventually exceed the constraints of a system. BRP can counter this increase in space by eliminating the need to store suboptimal paths of the game tree.

BRP shares some similarities to the earlier pruning algorithm *Regret-Based Pruning*, which has shown empirical evidence of improving the performance of CFR. In contrast, this paper proves that CFR converges faster when using BRP, because suboptimal paths in the game tree will only need to be traversed $O\big(\ln(T)\big)$ times over T iterations. We also prove that BRP uses asymptotically less space, while Regret-Based Pruning does not reduce the space needed to solve a game. Moreover, Best-Response Pruning easily generalizes to iterative algorithms beyond CFR such as Fictitious Play (Heinrich et al., 2015).

The magnitude of the gains in speed and space that BRP provides varies depending on the game. It is possible to construct games where BRP provides no benefit. However, if there are many suboptimal actions in the game—as is frequently the case in large games—BRP can speed up CFR by multiple orders of magnitude and require orders of magnitude less space. Our experiments show an order of magnitude space reduction already in medium-sized games, and a reduction factor increase with game size.

2. Background

In a two-player zero-sum imperfect-information extensive-form game there are two players, $\mathcal{P} = \{1, 2\}$. Let H be the set of all possible histories (nodes) in the game tree, represented as a sequence of actions. The actions available in a history is $A(h)$ and the player who acts at that history is $P(h) \in \mathcal{P} \cup c$, where c denotes chance. Chance plays an action $a \in A(h)$ with a fixed probability. The history h' reached after action a in h is a child of h, represented by $h \cdot a = h'$, while h is the parent of h'. More generally, h' is an ancestor of h (and h is a descendant of h'), represented by $h' \sqsubset h$, if there exists a sequence of actions from h' to h. $Z \subseteq H$ are terminal histories. For each player $i \in \mathcal{P}$, there is a payoff function $u_i : Z \to \Re$ where $u_1 = -u_2$. Define $\Delta_i = \max_{z \in Z} u_i(z) - \min_{z \in Z} u_i(z)$ and $\Delta = \max_i \Delta_i$.

Imperfect information is represented by *information sets* for each player $i \in \mathcal{P}$ by a partition $\mathcal{I}_i$ of $h \in H : P(h) = i$. For any information set $I \in \mathcal{I}_i$, all histories $h, h' \in I$ are indistinguishable to player i, so $A(h) = A(h')$. $I(h)$ is the information set I where $h \in I$. $P(I)$ is the player i such that $I \in \mathcal{I}_i$. $A(I)$ is the set of actions such that for all $h \in I$, $A(I) = A(h)$. $|A_i| = \max_{I \in \mathcal{I}_i} |A(I)|$ and $|A| = \max_i |A_i|$. Define $U(I)$ to be the maximum payoff reachable from a history in I, and $L(I)$ to be the minimum. That is, $U(I) = \max_{z \in Z, h \in I : h \sqsubseteq z} u_{P(I)}(z)$ and $L(I) = \min_{z \in Z, h \in I : h \sqsubseteq z} u_{P(I)}(z)$. Define $\Delta(I) = U(I) - L(I)$ to be the range of payoffs reachable from a history in I. Similarly $U(I, a)$, $L(I, a)$, and $\Delta(I, a)$ are the maximum, minimum, and range of payoffs (respectively) reachable from a history in I after taking action a. Define $D(I, a)$ to be the set of information sets reachable by player $P(I)$ after taking action a. Formally, $I' \in D(I, a)$ if for some history $h \in I$ and $h' \in I'$, $h \cdot a \sqsubseteq h'$ and $P(I) = P(I')$.

A strategy $\sigma_i(I)$ is a probability vector over $A(I)$ for player i in information set I. The probability of a particular action a is denoted by $\sigma_i(I, a)$. Since all histories in an information set belonging to player i are indistinguishable, the strategies in each of them must be identical. That is, for all $h \in I$, $\sigma_i(h) = \sigma_i(I)$ and $\sigma_i(h, a) = \sigma_i(I, a)$. Define σ_i to be a probability vector for player i over all available strategies Σ_i in the game. A strategy profile σ is a tuple of strategies, one for each player. $u_i(\sigma_i, \sigma_{-i})$ is the expected payoff for player i if all players play according to the strategy profile $\langle \sigma_i, \sigma_{-i} \rangle$. If a series of strategies are played over T iterations, then $\bar{\sigma}_i^T = \frac{\sum_{t \in T} \sigma_i^t}{T}$.

$\pi^\sigma(h) = \Pi_{h' \to a \sqsubseteq h} \sigma_{P(h')}(h', a)$ is the joint probability of reaching h if all players play according to σ. $\pi_i^\sigma(h)$ is the contribution of player i to this probability (that is, the probability of reaching h if all players other than i, and chance, always chose actions leading to h). $\pi_{-i}^\sigma(h)$ is the contribution of all players other than i, and chance. $\pi^\sigma(h, h')$ is the

probability of reaching h' given that h has been reached, and 0 if $h \not\sqsubset h'$. In a *perfect-recall* game, $\forall h, h' \in I \in \mathcal{I}_i$, $\pi_i(h) = \pi_i(h')$. In this paper we focus on perfect-recall games. Therefore, for $i = P(I)$ define $\pi_i(I) = \pi_i(h)$ for $h \in I$. Moreover, $I' \sqsubset I$ if for some $h' \in I'$ and some $h \in I$, $h' \sqsubset h$. Similarly, $I' \cdot a \sqsubset I$ if $h' \cdot a \sqsubset h$. The average strategy $\bar{\sigma}_i^T(I)$ for an information set I is defined as $\bar{\sigma}_i^T(I) = \frac{\sum_{t \in T} \pi_i^{\sigma_i^t}(I)\sigma_i^t(I)}{\sum_{t \in T} \pi_i^{\sigma^t}(I)}$.

A *best response* to σ_{-i} is a strategy σ_i^* such that $u_i(\sigma_i^*, \sigma_{-i}) = \max_{\sigma_i' \in \Sigma_i} u_i(\sigma_i', \sigma_{-i})$. A *Nash equilibrium* σ^* is a strategy profile where every player plays a best response: $\forall i$, $u_i(\sigma_i^*, \sigma_{-i}^*) = \max_{\sigma_i' \in \Sigma_i} u_i(\sigma_i', \sigma_{-i}^*)$. A *Nash equilibrium strategy* for player i as a strategy σ_i that is part of any Nash equilibrium. In two-player zero-sum games, if σ_i and σ_{-i} are both Nash equilibrium strategies, then $\langle \sigma_i, \sigma_{-i} \rangle$ is a Nash equilibrium. An *ϵ-equilibrium* as a strategy profile σ^* such that $\forall i$, $u_i(\sigma_i^*, \sigma_{-i}^*) + \epsilon \geq \max_{\sigma_i' \in \Sigma_i} u_i(\sigma_i', \sigma_{-i}^*)$.

2.1. Counterfactual Regret Minimization

Counterfactual Regret Minimization (CFR) is a popular algorithm for extensive-form games in which the strategy vector for each information set is determined according to a regret-minimization algorithm (Zinkevich et al., 2007). We use *regret matching (RM)* (Hart & Mas-Colell, 2000) as the regret-minimization algorithm, but the material presented in this paper also applies to other regret minimizing algorithms such as Hedge (Brown et al., 2017).

The analysis of CFR makes frequent use of *counterfactual value*. Informally, this is the expected utility of an information set given that player i tries to reach it. For player i at information set I given a strategy profile σ, this is defined as

$$v^\sigma(I) = \sum_{h \in I} \left(\pi_{-i}^\sigma(h) \sum_{z \in Z} \left(\pi^\sigma(h, z) u_i(z) \right) \right) \quad (1)$$

The counterfactual value of an action a is

$$v^\sigma(I, a) = \sum_{h \in I} \left(\pi_{-i}^\sigma(h) \sum_{z \in Z} \left(\pi^\sigma(h \cdot a, z) u_i(z) \right) \right) \quad (2)$$

A *counterfactual best response* (Moravcik et al., 2016) (CBR) is a strategy similar to a best response, except that it maximizes counterfactual value even at information sets that it does not reach due to its earlier actions. Specifically, a counterfactual best response to σ_{-i} is a strategy $CBR(\sigma_{-i})$ such that if $CBR(\sigma_{-i})(I, a) > 0$ then $v^{\langle CBR(\sigma_{-i}), \sigma_{-i} \rangle}(I, a) = \max_{a'} v^{\langle CBR(\sigma_{-i}), \sigma_{-i} \rangle}(I, a')$. The *counterfactual best response value* $CBV^{\sigma_{-i}}(I)$ is similar to counterfactual value, except that player $i = P(I)$ plays according to a CBR to σ_{-i}. Formally, $CBV^{\sigma_{-i}}(I) = v^{\langle CBR_i(\sigma_{-i}), \sigma_{-i} \rangle}(I)$.

Let σ^t be the strategy profile used on iteration t. The *instantaneous regret* on iteration t for action a in information set I is $r^t(I, a) = v^{\sigma^t}(I, a) - v^{\sigma^t}(I)$ and the *regret* for action a in I on iteration T is $R^T(I, a) = \sum_{t \in T} r^t(I, a)$. Additionally, $R_+^T(I, a) = \max\{R^T(I, a), 0\}$ and $R^T(I) = \max_a\{R_+^T(I, a)\}$. Regret for player i in the entire game is $R_i^T = \max_{\sigma_i' \in \Sigma_i} \sum_{t \in T} \left(u_i(\sigma_i', \sigma_{-i}^t) - u_i(\sigma_i^t, \sigma_{-i}^t) \right)$.

In regret matching, a player picks a distribution over actions in an information set in proportion to the positive regret on those actions. Formally, on each iteration $T + 1$, player i selects actions $a \in A(I)$ according to probabilities

$$\sigma^{T+1}(I, a) = \begin{cases} \frac{R_+^T(I, a)}{\sum_{a' \in A(I)} R_+^T(I, a')}, & \text{if } \sum_{a'} R_+^T(I, a') > 0 \\ \frac{1}{|A(I)|}, & \text{otherwise} \end{cases} \quad (3)$$

If a player plays according to RM on every iteration then on iteration T, $R^T(I) \leq \Delta(I)\sqrt{|A(I)|}\sqrt{T}$.

If a player plays according to CFR in every iteration then $R_i^T \leq \sum_{I \in \mathcal{I}_i} R^T(I)$. So, as $T \to \infty$, $\frac{R_i^T}{T} \to 0$. In two-player zero-sum games, if both players' average regret $\frac{R_i^T}{T} \leq \epsilon$, their average strategies $\langle \bar{\sigma}_1^T, \bar{\sigma}_2^T \rangle$ form a 2ϵ-equilibrium (Waugh et al., 2009). Thus, CFR constitutes an anytime algorithm for finding an ϵ-Nash equilibrium in zero-sum games.

2.2. Prior Approaches to Pruning

This section reviews forms of pruning that allow parts of the game tree to be skipped in CFR. In vanilla CFR, the entire game tree is traversed separately for each player history-by-history. On each traversal, the regret for each action of a history's information set is updated based on the expected value for that action on that iteration, weighed by the probability of opponents taking actions to reach the history (that is, weighed by $\pi_{-i}^{\sigma^t}(h)$). However, if a history h is reached on iteration t in which $\pi_{-i}^{\sigma^t}(h) = 0$, then from (1) and (2) the strategy at h contributes nothing on iteration t to the regret of $I(h)$ (or to the information sets above it). Moreover, any history that would be reached beyond h would also contribute nothing to its information set's regret because $\pi_{-i}^{\sigma^t}(h') = 0$ for every history h' where $h \sqsubset h'$ and $P(h') = P(h)$. Thus, when traversing the game tree for player i, there is no need to traverse beyond any history h when $\pi_{-i}^{\sigma^t}(h) = 0$. The benefit of this form of pruning, which we refer to as *partial pruning*, varies depending on the game, but empirical results show a factor of 30 improvement in some games (Lanctot et al., 2009).

While partial pruning allows one to prune paths that an *opponent* reaches with zero probability, Regret-Based Pruning allows one to also prune paths that the *traverser* reaches

with zero probability (Brown & Sandholm, 2015a). However, this pruning is necessarily temporary. Consider an action $a \in A(I)$ such that $\sigma^t(I,a) = 0$, and assume that it is known action a will not be played with positive probability until some far-future iteration t' (in RM, this would be the case if $R^t(I,a) \ll 0$). Since action a is played with zero probability on iteration t, so from (1) the strategy played and reward received following action a (that is, in $D(I,a)$) will not contribute to the regret for any information set preceding action a on iteration t. In fact, what happens in $D(I,a)$ has no bearing on the rest of the game tree until iteration t' is reached. So one could, in theory, "procrastinate" in deciding what happened beyond action a on iteration t, $t+1$, ..., $t'-1$ until iteration t'.

However, upon reaching iteration t', rather than individually making up the $t'-t$ iterations over $D(I,a)$, one can instead do a *single* iteration, playing against the average of the opponents' strategies in the $t'-t$ iterations that were missed, and declare that strategy was played on all the $t'-t$ iterations. This accomplishes the work of the $t'-t$ iterations in a single traversal. Moreover, since player i never plays action a with positive probability between iterations t and t', that means every *other* player can apply partial pruning on that part of the game tree for iterations $t'-t$, and skip it completely. This, in turn, means that player i has free rein to play whatever they want in $D(I,a)$ without affecting the regrets of the other players. In light of that, and of the fact that player i gets to decide what is played in $D(I,a)$ after knowing what the other players have played, player i might as well play a strategy that ensures zero regret for all information sets $I' \in D(I,a)$ in the iterations t to t'. A CBR to the average of the opponent strategies on the $t'-t$ iterations would qualify as such a zero-regret strategy.

Regret-Based Pruning only allows a player to skip traversing $D(I,a)$ for as long as $\sigma^t(I,a) = 0$. Thus, in RM, if $R^{t_0}(I,a) < 0$, we can prune the game tree beyond action a from iteration t_0 until iteration t_1 so long as $\sum_{t=1}^{t_0} v^{\sigma^t}(I,a) + \sum_{t=t_0+1}^{t_1} \pi_{-i}^{\sigma^t}(I) U(I,a) \leq \sum_{t=1}^{t_1} v^{\sigma^t}(I)$.

3. Best-Response Pruning

This section describes the behavior of BRP. In particular we focus on the case where BRP is applied to the most popular family of iterative algorithms, CFR. BRP begins pruning an action in an information set whenever playing perfectly beyond that action against the opponent's average strategy (that is, playing a CBR) still does worse than what has been achieved in the iterations played so far (that is, $\sum_{t=1}^{T} v^{\sigma^t}(I)$). Pruning continues for the minimum number of iterations it could take for the opponent's average strategy to change sufficiently such that the pruning starting condition (that is, playing a CBR beyond the action against the opponent's average strategy does worse than what has been achieved in the iterations so far) no longer holds. When pruning ends, BRP calculates a CBR in the pruned branch against the opponent's average strategy over all iterations played so far, and sets regret in the pruned branch as if that CBR strategy were played on *every* iteration played in the game so far—even those that were played before pruning began.

While using a CBR works correctly when applying BRP to CFR, it is also sound to choose a strategy that is *almost* a CBR (formalized later in this section), as long as that strategy ensures $\sum_{a \in A(I)} \left(R_+^T(I,a)\right)^2 \leq \left(\Delta(I)\right)^2 |A(I)|T$. In practice, this means that the strategy is close to a CBR, and approaches a CBR as $T \to \infty$. We now present the theory to show that such a near-CBR can be used. However, in practice CFR converges much faster than the theoretical bound, so the potential function is typically far lower than the theoretical bound. Thus, while choosing a near-CBR rather than an exact CBR may allow for slightly longer pruning according to the theory, it may actually result in worse performance. All of the theoretical results presented in this paper, including the improved convergence bound as well as the lower space requirements, still hold if only a CBR is used, and our experiments use a CBR. Nevertheless, clever algorithms for deciding on a near-CBR may lead to even better performance in practice.

We define a strategy $\beta(\sigma_{-i}, T)$ as a *T-near counterfactual best response* (T-near CBR) to σ_{-i} if for all I belonging to player i

$$\sum_{a \in A(I)} \left(v^{\langle \beta(\sigma_{-i},T), \sigma_{-i} \rangle}(I,a) - v^{\langle \beta(\sigma_{-i},T), \sigma_{-i} \rangle}(I)\right)_+^2 \leq \frac{x_I^T}{T^2} \tag{4}$$

where x_I^T can be any value in the range $0 \leq x_I^T \leq \left(\Delta(I)\right)^2 |A(I)|T$. If $x_I^T = 0$, then a T-near CBR is always a CBR. The set of strategies that are T-near CBRs to σ_{-i} is represented as $\Sigma^\beta(\sigma_{-i}, T)$. We also define the *$T$-near counterfactual best response value* as $\psi^{\sigma_{-i},T}(I,a) = \min_{\sigma_i' \in \Sigma^\beta(\sigma_{-i},T)} v^{\langle \sigma_i', \sigma_{-i} \rangle}(I,a)$ and $\psi^{\sigma_{-i},T}(I) = \min_{\sigma_i' \in \Sigma^\beta(\sigma_{-i},T)} v^{\langle \sigma_i', \sigma_{-i} \rangle}(I)$.

When applying BRP to CFR, an action is pruned only if it would still have negative regret had a T-near CBR against the opponent's average strategy been played on every iteration. Specifically, on iteration T of CFR with RM, if

$$T\left(\psi^{\bar{\sigma}_{-i}^T,T}(I,a)\right) \leq \sum_{t=1}^{T} v^{\sigma^t}(I) \tag{5}$$

then $D(I,a)$ can be pruned for

$$T' = \frac{\sum_{t=1}^{T} v^{\sigma^t}(I) - \psi^{\bar{\sigma}^T_{-i},T}(I,a)}{U(I,a) - L(I)} \tag{6}$$

iterations. After those T' iterations are over, we calculate a $T+T'$-near CBR in $D(I,a)$ to the opponent's average strategy and set regret as if that $T+T'$-near CBR had been played on every iteration. Specifically, for each $t \leq T+T'$ we set[2] $v^{\sigma^t}(I,a) = \psi^{\bar{\sigma}^{T+T'}_{-i},T+T'}(I,a)$ so that

$$R^{T+T'}(I,a) = \big(T+T'\big)\big(\psi^{\bar{\sigma}^{T+T'}_{-i},T+T'}(I,a)\big) - \sum_{t=1}^{T+T'} v^{\sigma^t}(I) \tag{7}$$

and for every information set $I' \in D(I,a)$ we set $v^{\sigma^t}(I',a') = \psi^{\bar{\sigma}^{T+T'}_{-i},T+T'}(I',a')$ and $v^{\sigma^t}(I') = \psi^{\bar{\sigma}^{T+T'}_{-i},T+T'}(I')$ so that

$$R^{T+T'}(I',a') = \big(T+T'\big)\big(\psi^{\bar{\sigma}^{T+T'}_{-i},T+T'}(I',a') - \psi^{\bar{\sigma}^{T+T'}_{-i},T+T'}(I')\big) \tag{8}$$

Theorem 1 proves that if (5) holds for some action, then the action can be pruned for T' iterations, where T' is defined in (6). The same theorem holds if one replaces the T-near counterfactual best response values with exact counterfactual best response values. The proof for Theorem 1 draws from recent work on warm starting CFR using only an average strategy profile (Brown & Sandholm, 2016). Essentially, we warm start regrets in the pruned branch using only the average strategy of the opponent and knowledge of T.

Theorem 1. *Assume T iterations of CFR with RM have been played in a two-player zero-sum game and assume $T\big(\psi^{\bar{\sigma}^T_{-i},T}(I,a)\big) \leq \sum_{t=1}^{T} v^{\sigma^t}(I)$ where $P(I) = i$. Let $T' = \lfloor \frac{\sum_{t=1}^{T} v^{\sigma^t}(I) - T\left(\psi^{\bar{\sigma}^T_{-i},T}(I,a)\right)}{U(I,a)-L(I)} \rfloor$. If both players play according to CFR with RM for the next T' iterations in all information sets $I'' \notin D(I,a)$ except that $\sigma(I,a)$ is set to zero and $\sigma(I)$ is renormalized, then $(T+T')\big(\psi^{\bar{\sigma}^{T+T'}_{-i},T+T'}(I,a)\big) \leq \sum_{t=1}^{T+T'} v^{\sigma^t}(I)$. Moreover, if one then sets $v^{\sigma^t}(I,a) = \psi^{\bar{\sigma}^{T+T'}_{-i},T+T'}(I,a)$ for each $t \leq T+T'$ and $v^{\sigma^t}(I',a') = \psi^{\bar{\sigma}^{T+T'}_{-i},T+T'}(I',a')$ for each $I' \in D(I,a)$, then after T'' additional iterations of CFR with RM, the bound on exploitability of $\bar{\sigma}^{T+T'+T''}$ is no worse than having played $T+T'+T''$ iterations of CFR with RM without BRP.*

In practice, rather than check whether (5) is met for an action on every iteration, one could only check actions that have very negative regret, and do a check only once every several iterations. This would still be safe and would save some computational cost of the checks, but would lead to less pruning.

Similar to Regret-Based Pruning, the duration of pruning in BRP can be increased by giving up knowledge beforehand of exactly how many iterations can be skipped. From (2) and (1) we see that $r^T(I,a) \leq \pi^{\sigma^t}_{-i}(I)\big(U(I,a) - L(I)\big)$. Thus, if $\pi^{\sigma^t}_{-i}(I)$ is very low, then (5) would continue to hold for more iterations than (6) guarantees. Specifically, we can prune $D(I,a)$ from iteration t_0 until iteration t_1 as long as

$$t_0\big(\psi^{\bar{\sigma}^{t_0}_{-i},t_0}(I,a)\big) + \sum_{t=t_0+1}^{t_1} \pi^{\sigma^t}_{-i}(I)U(I,a) \leq \sum_{t=1}^{t_1} v^{\sigma^t}(I) \tag{9}$$

3.1. Best-Response Pruning Requires Less Space

A key advantage of BRP is that setting the new regrets according to (7) and (8) requires no knowledge of what the regrets were before pruning began. Thus, once pruning begins, all the regrets in $D(I,a)$ can be discarded and the space that was allocated to storing the regret can be freed. That space need only be re-allocated once (9) ceases to hold *and* we cannot immediately begin pruning again (that is, (5) does not hold). Theorem 2 proves that for any information set I and action $a \in A(I)$ that is not part of a best response to a Nash equilibrium, there is an iteration $T_{I,a}$ such that for all $T \geq T_{I,a}$, action a in information set I (and its descendants) can be pruned.[3] Thus, once this $T_{I,a}$ is reached, it will never be necessary to allocate space for regret in $D(I,a)$ again.

Theorem 2. *In a two-player zero-sum game, if for every opponent Nash equilibrium strategy $\sigma^*_{-P(I)}$, $CBV^{\sigma^*_{-P(I)}}(I,a) < CBV^{\sigma^*_{-P(I)}}(I)$, then there exists a $T_{I,a}$ and $\delta_{I,a} > 0$ such that after $T \geq T_{I,a}$ iterations of CFR, $CBV^{\bar{\sigma}^T_{-i}}(I,a) - \frac{\sum_{t=1}^{T} v^{\sigma^t}(I)}{T} \leq -\delta_{I,a}$.*

While such a constant $T_{I,a}$ exists for any suboptimal action, BRP cannot determine whether or when $T_{I,a}$ is reached. Thus, it is still necessary to check whether (5) is satisfied whenever (9) no longer holds, and to recalculate how much longer $D(I,a)$ can safely be pruned. This requires the algorithm to periodically calculate a best response (or near-best response) in $D(I,a)$. However, this (near-)best response calculation does not require knowledge of regret in $D(I,a)$,

[2] In practice, only the sums $\sum_{t=1}^{T} v^{\sigma^t}(I)$ and either $\sum_{t=1}^{T} v^{\sigma^t}(I,a)$ or $R^T(I,a)$ are stored.

[3] If CFR converges to a *particular* Nash equilibrium, then this condition could be broadened to any information set I and action $a \in A(I)$ that is not a best response to *that particular* Nash equilibrium. While empirically CFR does appear to always converge to a particular Nash equilibrium, there is no known proof that it always does so.

so it is still never necessary to store regret after iteration $T_{I,a}$ is reached.

While it is possible to discard regrets in $D(I,a)$ without penalty once pruning begins, regret is only half the space requirement of CFR. Every information set I also stores a sum of the strategies played $\sum_{t=1}^{T}\left(\pi_i^{\sigma^t}(I)\sigma^t(I)\right)$ which is normalized once CFR ends in order to calculate $\bar{\sigma}^T(I)$. Fortunately, if action a in information set I is pruned for long enough, then the stored cumulative strategy in $D(I,a)$ can also be discarded at the cost of a small increase in the distance of the final average strategy from a Nash equilibrium. Specifically, if $\pi_i^{\bar{\sigma}^T}(I,a) \leq \frac{C}{\sqrt{T}}$, where C is some constant, then setting $\bar{\sigma}^T(I,a) = 0$ and renormalizing $\bar{\sigma}^T(I)$, and setting $\bar{\sigma}^T(I',a') = 0$ for $I' \in D(I,a)$, can result in at most $\frac{C|I|\Delta}{\sqrt{T}}$ higher exploitability for the whole strategy $\bar{\sigma}^T$. Since CFR only guarantees that $\bar{\sigma}^T$ is a $\frac{2|\mathcal{I}|\Delta\sqrt{|A|}}{\sqrt{T}}$-Nash equilibrium anyway, $\frac{C|I|\Delta}{\sqrt{T}}$ is only a constant factor of the bound. If an action is pruned from T' to T, then $\sum_{t=1}^{T}\left(\pi_i^{\sigma^t}(I)\sigma^t(I,a)\right) \leq \frac{T'}{T}$. Thus, if an action is pruned for long enough, then eventually $\sum_{t=1}^{T}\left(\pi_i^{\sigma^t}(I)\sigma^t(I,a)\right) \leq \frac{C}{\sqrt{T}}$ for any C, so $\sum_{t=1}^{T}\left(\pi_i^{\sigma^t}(I)\sigma^t(I,a)\right)$ could be set to zero (as well as all descendants of $I \cdot a$), while suffering at most a constant factor increase in exploitability. As more iterations are played, this penalty will continue to decrease and eventually be negligible. The constant C can be set by the user: a higher C allows the average strategy to be discarded sooner, while a lower C reduces the potential penalty in exploitability.

We define $\mathcal{I}_S$ as the set of information sets that are not guaranteed to be asymptotically pruned by Theorem 2. Specifically, $I \in \mathcal{I}_S$ if $I \notin D(I',a')$ for some I' and $a' \in A(I')$ such that for every opponent Nash equilibrium strategy $\sigma^*_{-P(I')}$, $CBV^{\sigma^*_{-P(I')}}(I',a') < CBV^{\sigma^*_{-P(I')}}(I')$. Theorem 2 implies the following.

Corollary 1. *In a two-player zero-sum game with some threshold on the average strategy $\frac{C}{\sqrt{T}}$ for $C > 0$, after a finite number of iterations CFR with BRP requires only $O\left(|\mathcal{I}_S||A|\right)$ space.*

Using a threshold of $\frac{C}{T}$ rather than $\frac{C}{\sqrt{T}}$ does not change the theoretical properties of the corollary, and may lead to faster convergence in some situations, but it may also result in a slower reduction in the space used by the algorithm (though the asymptotic space used is identical). In particular, if BRP can be extended to first-order methods that converge to an ϵ-Nash equilibrium in $O(\frac{1}{\epsilon})$ iterations rather than $O(\frac{1}{\epsilon^2})$ iterations, such as the Excessive Gap Technique (Hoda et al., 2010; Kroer et al., 2017), then a threshold of $\frac{C}{T}$ may be more appropriate when those algorithms are used. A threshold of $\frac{C}{T}$ may also be preferable when using an algorithm which empirically converges to an ϵ-Nash equilibrium in faster than $O(\frac{1}{\epsilon^2})$ iterations, such as CFR+ on some games.

3.2. Best-Response Pruning Converges Faster

We now prove that BRP in CFR speeds up convergence to an ϵ-Nash equilibrium. Section 3 proved that CFR with BRP converges in the same number of iterations as CFR alone. In this section, we prove that BRP allows each iteration to be traversed more quickly. Specifically, if an action $a \in A(I)$ is not a CBR to a Nash equilibrium, then $D(I,a)$ need only be traversed $O(\ln(T))$ times over T iterations. Intuitively, as both players converge to a Nash equilibrium, actions that are not a counterfactual best response will eventually do worse than actions that are, so those suboptimal actions will accumulate increasing amounts of negative regret. This negative regret allows the action to be safely pruned for increasingly longer periods of time.

Specifically, let $S \subseteq H$ be the set of histories where $h \cdot a \in S$ if $h \in S$ and action a is part of some CBR to some Nash equilibrium. Formally, S contains $\emptyset$ and every history $h \cdot a$ such that $h \in S$ and $CBV^{\sigma^*_{-P(I)}}(I,a) = CBV^{\sigma^*_{-P(I)}}(I)$ for some Nash equilibrium σ^*.

Theorem 3. *In a two-player zero-sum game, if both players choose strategies according to CFR with BRP, then conducting T iterations requires only $O\left(|S|T + |H|\ln(T)\right)$ nodes to be traversed.*

The definition of S uses properties of the Nash equilibria of the game, and an action $a \in A(I)$ not in S is only guaranteed to be pruned by BRP after some $T_{I,a}$ is reached, which also depends on the Nash equilibria of the game. Since CFR converges to only an ϵ-Nash equilibrium, CFR cannot determine with certainty which nodes are in S or when $T_{I,a}$ is reached. Nevertheless, both S and $T_{I,a}$ are fixed properties of the game.

4. Experiments

We compare the convergence speed of BRP to Regret-Based Pruning, to only partial pruning, and to no pruning at all. We also show that BRP uses less space as as more iterations are conducted, unlike prior pruning algorithms. The experiments are conducted on Leduc Hold'em (Southey et al., 2005) and Leduc-5 (Brown & Sandholm, 2015a). Leduc Hold'em is a common benchmark in imperfect-information game solving because it is small enough to be solved but still strategically complex. In Leduc Hold'em, there is a deck consisting of six cards: two each of Jack, Queen, and King. There are two rounds. In the first round, each player places an ante of 1 chip in the pot and receives a single private card. A round of betting then takes place with a two-bet maximum, with Player 1 going first. A public shared card is then dealt face up and another round of betting takes place.

Again, Player 1 goes first, and there is a two-bet maximum. If one of the players has a pair with the public card, that players wins. Otherwise, the player with the higher card wins. The bet size in the first round is 2 chips, and 4 chips in the second round. Leduc-5 is like Leduc Hold'em but larger: there are 5 bet sizes to choose from. In the first round a player may bet 0.5, 1, 2, 4, or 8 chips, while in the second round a player may bet 1, 2, 4, 8, or 16 chips.

Nodes touched is a hardware and implementation-independent proxy for time which we use to measure performance of the various algorithms. Overhead costs are counted in nodes touched. CFR+ is a variant of CFR in which a floor on regret is set at zero and each iteration is weighted linearly in the average strategy (that is, iteration t is weighted by t) rather than each iteration being weighted equally. Since Regret-Based Pruning can only prune negative-regret actions, Regret-Based Pruning modifies the definition of CFR+ so that regret can be negative, but immediately jumps up to zero as soon as regret increases. BRP does not require this modification. Still, BRP also modifies the behavior of CFR+ because without pruning, CFR+ would put positive probability on an action as soon as its regret increases, while BRP waits until pruning is over. This is not, by itself, a problem. However, CFR+'s linear weighting of the average strategy is only guaranteed to converge to a Nash equilibrium if pruning does not occur. While both Regret-Based Pruning and BRP do well empirically with CFR+, the convergence is noisy. This noise can be reduced by using the lowest-exploitability average strategy profile found so far, which we do in the experiments.[4] BRP does not do as well empirically with the linear-averaging component of CFR+. Thus, for BRP we only measure performance using RM+ with CFR, which is the same as CFR+ but without linear averaging. CFR+ with and without linear averaging has the same theoretical performance as CFR, but CFR+ does better empirically (particularly with linear averaging).

Figure 1 and Figure 2 show the reduction in space needed to store the average strategy and regrets for BRP—for various values of the constant threshold C, where an action's probability is set to zero if it is reached with probability less than $\frac{C}{\sqrt{T}}$ in the average strategy, as we explained in Section 3.1. In both games, a threshold between 0.01 and 0.1 performed well in both space and number of iterations, with the lower thresholds converging somewhat faster and the higher thresholds reducing space faster. We also tested thresholds below 0.01, but the speed of convergence was essentially the same as when using 0.01. In Leduc, all variants resulted in a quick drop-off in space to about half the initial amount. In Leduc-5, a threshold of 0.1 resulted in about a factor of 7 reduction for both CFR with RM and CFR with RM+. This space reduction factor appears to continue to increase.

[4]Exploitability is no harder to compute than one iteration of CFR or CFR+. Snapshots are not plotted at every iteration but only after every 10,000,000 nodes touched—except for the first few snapshots.

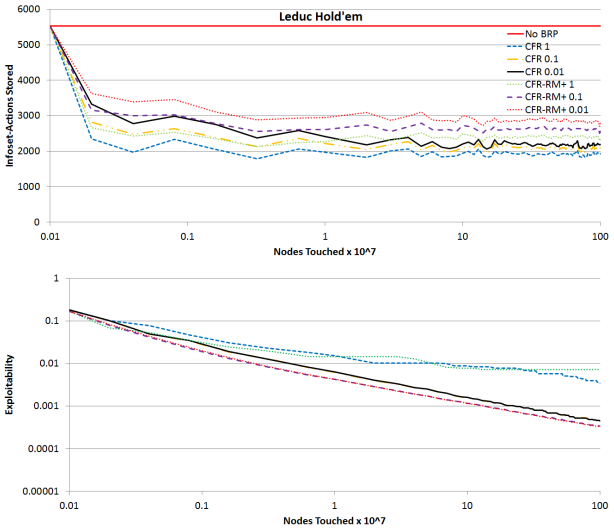

Figure 1. Convergence and space required for CFR using RM and RM+ with best-response pruning in Leduc Hold'em. The y-axis on the top graph is linear scale.

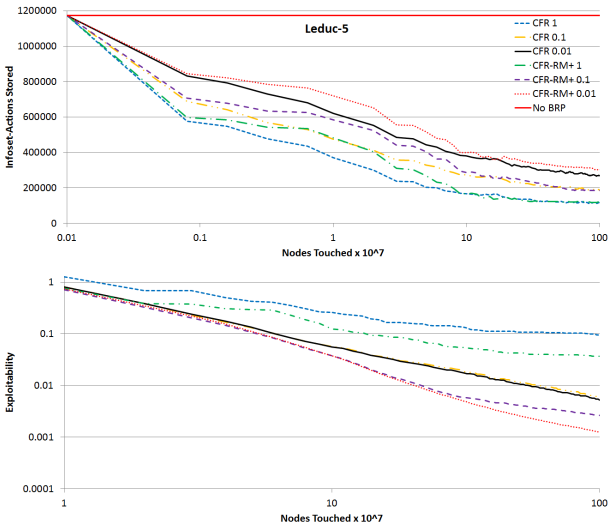

Figure 2. Convergence and space required for CFR using RM and RM+ with best-response pruning in Leduc-5. The y-axis on the top graph is linear scale.

Figure 3 and Figure 4 compare the convergence rates of BRP, Regret-Based Pruning, and only partial pruning for CFR with RM, CFR with RM+, and CFR+. In Leduc, BRP and Regret-Based Pruning perform comparably when added to CFR. Regret-Based Pruning with CFR+ does significantly better, while BRP with CFR using RM+ sees no improve-

ment over BRP with CFR. In Leduc-5, which is a far larger game, BRP outperforms Regret-Based Pruning by a factor of 2 when added to CFR. BRP with CFR using RM+ also performs comparably to Regret-Based Pruning with CFR+, while retaining theoretical guarantees and not suffering from noisy convergence.

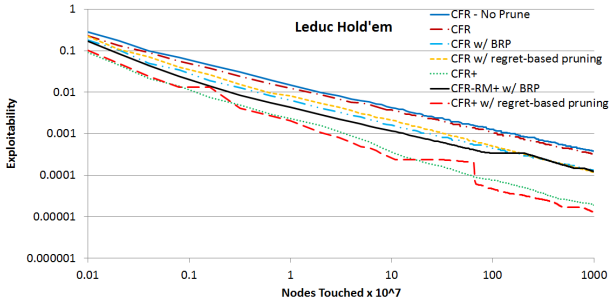

Figure 3. Convergence for partial pruning, regret-based pruning, and best-response pruning in Leduc. "CFR - No Prune" is CFR without any pruning.

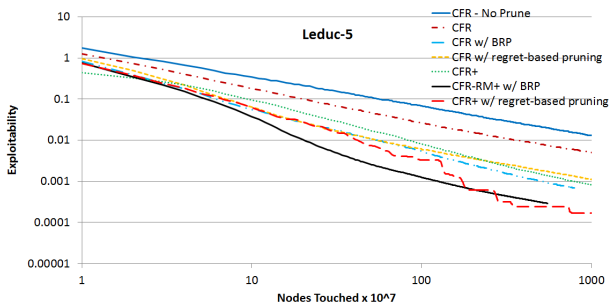

Figure 4. Convergence for partial pruning, regret-based pruning, and best-response pruning in Leduc-5. "CFR - No Prune" is CFR without any pruning.

5. Conclusions

We introduced BRP, a new form of pruning that provably reduces both the space needed to solve an imperfect-information game and the time needed to reach an ϵ-Nash equilibrium. This addresses both of the major bottlenecks in solving large imperfect-information games. Experimentally, BRP reduced the space needed to solve a game by a factor of 7, with the reduction factor increasing with game size. While the early iterations may still be slow and require the same amount of space as CFR without BRP, these early iterations can be skipped by warm starting CFR with an abstraction of the game (Brown & Sandholm, 2016). This paper focused on the theory of BRP when applied to CFR, the most popular algorithm for solving imperfect-information games. However, BRP can also be applied to Fictitious Play (Heinrich et al., 2015) and likely extends to other iterative algorithms as well (Hoda et al., 2010).

6. Acknowledgments

This material is based on work supported by the National Science Foundation under grant IIS-1617590 and the ARO under award W911NF-17-1-0082.

References

Bowling, Michael, Burch, Neil, Johanson, Michael, and Tammelin, Oskari. Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149, January 2015.

Brown, Noam and Sandholm, Tuomas. Regret-based pruning in extensive-form games. In *Advances in Neural Information Processing Systems*, pp. 1972–1980, 2015a.

Brown, Noam and Sandholm, Tuomas. Simultaneous abstraction and equilibrium finding in games. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2015b.

Brown, Noam and Sandholm, Tuomas. Strategy-based warm starting for regret minimization in games. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

Brown, Noam, Kroer, Christian, and Sandholm, Tuomas. Dynamic thresholding and pruning for regret minimization. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

Gilpin, Andrew and Sandholm, Tuomas. Lossless abstraction of imperfect information games. *Journal of the ACM*, 54(5), 2007. Early version 'Finding equilibria in large sequential games of imperfect information' appeared in the Proceedings of the ACM Conference on Electronic Commerce (EC), pages 160–169, 2006.

Hart, Sergiu and Mas-Colell, Andreu. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150, 2000.

Heinrich, Johannes, Lanctot, Marc, and Silver, David. Fictitious self-play in extensive-form games. In *International Conference on Machine Learning (ICML)*, pp. 805–813, 2015.

Hoda, Samid, Gilpin, Andrew, Peña, Javier, and Sandholm, Tuomas. Smoothing techniques for computing Nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2):494–512, 2010. Conference version appeared in WINE-07.

Kroer, Christian, Waugh, Kevin, Kılınç-Karzan, Fatma, and Sandholm, Tuomas. Faster first-order methods for extensive-form game solving. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, 2015.

Kroer, Christian, Waugh, Kevin, Kılınç-Karzan, Fatma, and Sandholm, Tuomas. Theoretical and practical advances on smoothing for extensive-form games. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, 2017.

Lanctot, Marc, Waugh, Kevin, Zinkevich, Martin, and Bowling, Michael. Monte Carlo sampling for regret minimization in extensive games. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 1078–1086, 2009.

Moravcik, Matej, Schmid, Martin, Ha, Karel, Hladik, Milan, and Gaukrodger, Stephen. Refining subgames in large imperfect information games. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

Nesterov, Yurii. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal of Optimization*, 16 (1):235–249, 2005.

Pays, François. An interior point approach to large games of incomplete information. In *AAAI Computer Poker Workshop*, 2014.

Schmid, Martin, Moravcik, Matej, and Hladik, Milan. Bounding the support size in extensive form games with imperfect information. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 784–790, 2014.

Southey, Finnegan, Bowling, Michael, Larson, Bryce, Piccione, Carmelo, Burch, Neil, Billings, Darse, and Rayner, Chris. Bayes' bluff: Opponent modelling in poker. In *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 550–558, July 2005.

Tammelin, Oskari, Burch, Neil, Johanson, Michael, and Bowling, Michael. Solving heads-up limit texas hold'em. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 645–652, 2015.

Waugh, Kevin, Schnizlein, David, Bowling, Michael, and Szafron, Duane. Abstraction pathologies in extensive games. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2009.

Zinkevich, Martin, Johanson, Michael, Bowling, Michael H, and Piccione, Carmelo. Regret minimization in games with incomplete information. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 1729–1736, 2007.