Efficient CFR for Imperfect Information Games with Instant Updates

Hui Li [1] **Kailiang Hu** [1] **Yuan Qi** [1] **Le Song** [1 2]

Abstract

Counterfactual regret minimization (CFR) is a framework of iterative algorithms and is empirically the fastest approach to solving large imperfect information games. However, for large games, the convergence speed of the state-of-the-art CFR is still the key limitation, especially in real-time applications. We propose a novel counterfactual regret minimization method with instant updates, which has a provably lower convergence bound and a provably tighter space complexity bound. We apply the proposed instant updates into many CFR variants on one Leduc Hold'em instance and five different subgame instances of Heads-Up No-Limit Texas Hold'em (HUNL) generated by DeepStack. The proposed method empirically achieves faster convergence rates than the state-of-the-art CFR. In subgame instances of HUNL, our method converges three times faster than the hybrid method used in DeepStack.

1. Introduction

In recent years, many remarkable advances have been made in addressing large perfect information games, such as Go (Silver et al., 2016; 2017). However, solving Imperfect Information Games (IIG) still remains a challenging problem. In IIGs, a player has only partial knowledge about her opponents before making a decision, so that she has to reason under the uncertainty about her opponents' information while exploiting the other players' uncertainty about herself. Thus, IIGs provide more realistic modeling than perfect information games for many real-world applications, such as trading, traffic routing, and public auction. The typical target of solving IIGs is to find a Nash equilibrium so that no player can unilaterally improve her reward.

To solve IIGs, many algorithms have been designed to approximately find Nash equilibrium. Linear programming with realization plan representation (Koller & Megiddo, 1992) has traditionally been used to solve perfect-recall constant-sum IIGs. Such representation is linear in the number of nodes in the game tree but usually requires inverting large matrix or other extremely expensive operation. Many iterative techniques have been proposed as an alternative to linear programming methods, such as gradient-based algorithm (Gilpin et al., 2007), excessive gap technique (Kroer et al., 2015) and regret minimization method (Gordon, 2007; Zinkevich et al., 2007). The widely used approaches for solving large IIGs are the CFR variants (Zinkevich et al., 2007; Lanctot et al., 2009; Tammelin, 2014; Brown & Sandholm, 2018; Schmid et al., 2018; Li et al., 2018), which minimize the overall counterfactual regret so that the average strategies converge to Nash equilibria. Zinkevich et al. (2007) uses CFR to solve the abstracted limit Texas Hold'em with 10^{12} states, which is two orders of magnitude larger than previous methods. To obtain a faster convergence, Tammelin et al. (2015); Tammelin (2014) propose CFR+ and ultimately solve Heads-Up Limit Texas Holdem (HUL) with CFR+ by 4800 CPUs and running for 68 days. Note that, this game has over 10^{14} information sets and has been a challenging problem for artificial intelligence over 10 years (Michael Bowling, 2015). Although great breakthroughs have been made, still Heads-Up No-Limit (HUNL) Texas Hold'em still remains an open question, which has more than 6×10^{161} information sets (Johanson, 2013) and is much more difficult than HUL. Recently, Libratus (Brown & Sandholm, 2017) and DeepStack (Moravcik et al., 2017) are developed to solve the abstracted versions of HUNL using CFR variants and continue resolving techniques. Because the agents have to solve the subgames online using CFR variants, to timely return the computed strategy profile, they have to reduce the size of subgame by abstraction technique.

To make it possible to solve larger IIGs with more-refined abstracted actions, a more efficient method is quite important and necessary. Brown & Sandholm (2018) propose a faster regret minimization method — DCFR — by discounting both positive and negative cumulative regret. This work won the honorable mention in AAAI 2019 and can achieve the fastest convergence rate on many subgame instances of HUNL empirically. In the experiment, we will compare our method against this method.

In this paper, we propose a more efficient counterfactual regret minimization method with instant updates technique. We prove that our method has a lower convergence bound

[1]Ant Financial, China [2]Georgia Institute of Technology, USA. Correspondence to: Hui Li <lihuiknight@google.com, ken.lh@antfin.com>.

Real-World Sequential Decision Making workshop at the 36^{th} International Conference on Machine Learning, 2019.

under the same proved computation memory constraint. More importantly, many popular and state-of-the-art CFR variants, such as original CFR (Zinkevich et al., 2007), CFR+ (Tammelin, 2014; Michael Bowling, 2015) and DCFR (Brown & Sandholm, 2018), can benefit from the proposed instant updates. We test our method on Leduc Hold'em and five different HUNL subgames generated by DeepStack, the experiment results show that the proposed instant updates technique makes significant improvements against CFR, CFR+, and DCFR. In addition, we also prove that the weighted average strategy by skipping previous iterations can approach an approximate Nash equilibrium. In the subgame instance of HUNL, the improved method converges three times faster than the hybrid method used in DeepStack.

2. Background and Notation

2.1. Notations in Extensive-Form Game

We define the components of an extensive-form IIG following (Osborne & Rubinstein, 1994; Li et al., 2018). $N=\{0,1,...,n-1\}$ is a finite set and each member refers to a player. We use h_i^v refer to the *hidden variable* of player i in imperfect information game, which is unobserved by the opponents. Each member h of H denotes a possible *history* (or state). For player i, h_{-i}^v refers to the opponent's hidden variables. If h_j is a prefix of h, we can denote them by $h_j \sqsubseteq h$. Z denotes the set of terminal histories and any member $z \in Z$ is not a prefix of any other sequences. A **player function** P assigns a member of $N \cup \{c\}$ to each non-terminal history, where c denotes the chance player. In practice, we usually define $c=-1$. $P(h)$ is the player who takes actions after history h. $A(h)=\{a : ha \in H\}$ is the set of available actions after non-terminal history $h \in H \setminus Z$. $\mathcal{I}_i$ of a history $\{h \in H : P(h)=i\}$ is an **information partition** of player i. A set $I_i \in \mathcal{I}_i$ is an **information set** of player i and $I_i(h)$ refers to information set I_i at state h. For $I_i \in \mathcal{I}_i$, we have $A(I_i)=A(h)$ and $P(I_i)=P(h)$. For each player $i \in N$, the utility function $u_i(z)$ defines the payoff of the terminal state z. If all players in one game can recall their previous actions and the corresponding infosets, we call it a **perfect-recall** game.

2.2. Definition of Strategy and Nash equilibrium

For play $i \in N$, the **strategy** $\sigma_i(I_i)$ in an extensive-form game assigns an action distribution over $A(I_i)$ to information set I_i. A **strategy profile** $\sigma=\{\sigma_i|\sigma_i \in \Sigma_i, i \in N\}$ is a collection of strategies for all players, where Σ_i is the set of all possible strategy profiles for player i. σ_{-i} refers to all strategies in σ expect σ_i. $\sigma_i(I_i)$ is the strategy of information set I_i. $\sigma_i(a|h)$ denotes the probability of action a taken by player $i \in N \cup \{c\}$ at state h. In imperfect information game, $\forall h_1 \in I_i$ and $\forall h_2 \in I_i$, we have $I_i=I_i(h_1)=I_i(h_2)$, $\sigma_i(I_i)=\sigma_i(h_1)=\sigma_i(h_2)$, $\sigma_i(a|I_i)=\sigma_i(a|h_1)=\sigma_i(a|h_2)$. For iterative learning method such as CFR, σ^t refers to the strategy profile at t-th iteration. The **state reach probability** of history h is denoted by $\pi^\sigma(h)$ if players take actions according to σ. For an empty sequence, $\pi^\sigma(a^0)=\pi^\sigma(\emptyset)=1$. The reach probability can be decomposed into $\pi^\sigma(h)=\prod_{i\in N\cup\{c\}}\pi_i^\sigma(h)=\pi_i^\sigma(h)\pi_{-i}^\sigma(h)$, where π_i^σ is the product of player $i's$ contribution and π_{-i}^σ is the product of all players' contribution except player i. The **information set reach probability** of I_i is defined by $\pi^\sigma(I_i)=\sum_{h\in I_i}\pi^\sigma(h)$. For player i, the **expected game utility** of a strategy profile σ is the expected payoff of all possible terminal nodes, *i.e.*, $u_i^\sigma=\sum_{z\in Z}\pi^\sigma(z)u_i(z)$. Given a fixed strategy profile σ_{-i}, any strategy $\sigma_i^*=\operatorname{argmax}_{\sigma_i'\in\Sigma_i} u_i^{(\sigma_i',\sigma_{-i})}$ of player i that achieves optimal payoff against π_{-i}^σ is a **best response**. An ϵ**-Nash equilibrium** is an approximation of a Nash equilibrium, whose strategy profile σ^* satisfies: $\forall i \in N$, $u_i^{(\sigma_i^*,\sigma_{-i})}+\epsilon \geq \max_{\sigma_i'\in\Sigma_i} u_i^{(\sigma_i',\sigma_{-i})}$. Exploitability of a strategy σ_i is defined by $\epsilon_i(\sigma_i)=u_i^{\sigma^*}-u_i^{(\sigma_i,\sigma_{-i}^*)}$. If the players alternate their positions in two-player zero-sum IIG, the value of a pair of games is zeros, *i.e.*, $u_0^{\sigma^*}+u_1^{\sigma^*}=0$. Therefore, we can define the **exploitability** of a strategy profile σ by $\epsilon(\sigma)=\frac{u_1^{(\sigma_0,\sigma_1^*)}+u_0^{(\sigma_0^*,\sigma_1)}}{2}$.

3. Method and Theory

In this section, we will present a novel regret minimization method with an efficient instant updates technique. Then we give the theoretical bound for this novel method. After that, we present another regret minimization method with skipping mechanism and prove its bound. At last, we talk about several hybrid methods of current CFR variants and the proposed instant updates.

3.1. Instant Counterfactual Regret Minimization

CFR variants (Zinkevich et al., 2007; Lanctot et al., 2009; Brown & Sandholm, 2017; Moravcik et al., 2017) update counterfactual value recursively along the game tree and minimize the overall regret. Our method also minimizes the overall regret. Different from original CFR variants, we define a novel instant counterfactual value recursively as follows.

Given the children's *instant counterfactual value* $s_i^{\sigma^t}(a|I_i)$ of information set I_i, its *dummy counterfactual value* is defined by

$$\hat{s}_i^{\sigma^t}(I_i)=\sum_{a\in A(I_i)}\sigma_i^t(a|I_i)s_i^{\sigma^t}(a|I_i). \tag{1}$$

Specifically, the leaf nodes' instant counterfactual values are the same as their utility values. Then the *instant regret* of taking action a at information set I_i will be

$$\hat{q}_i^{\sigma^t}(a|I_i)=s_i^{\sigma^t}(a|I_i)-\hat{s}_i^{\sigma^t}(I_i). \tag{2}$$

The *cumulative instant regret* is the rectified summation of total instant regret, which is defined by

$$Q_i^t(a|I_i)=max(Q_i^{t-1}(a|I_i)+\hat{q}_i^{\sigma^t}(a|I_i),0) \tag{3}$$

Then we update the behavior strategy $\sigma_i^{t+1}(a|I_i)$ by

$$\sigma_i^{T+1}(a|I_i)=\begin{cases}\frac{Q_i^t(a|I_i)}{\sum_{a\in A(I_i)}Q_i^t(a|I_i)} & \text{if} \sum_{a\in A(I_i)}Q_i^t(a|I_i)>0\\ \frac{1}{|A(I_i)|} & \text{otherwise.}\end{cases} \tag{4}$$

After that, instant counterfactual value $s_i^{\sigma^t}(I_i)$ of information set I_i is defined by

$$s_i^{\sigma^t}(I_i)=\sum_{a\in A(I_i)}\sigma_i^{t+1}(a|I_i)s_i^{\sigma^t}(a|I_i). \tag{5}$$

Now, we finish the recursive definition of instant counterfactual value and cumulative instant regret. Note that, the definition of counterfactual value in our method is different from the previous CFR variants (Zinkevich et al., 2007; Tammelin, 2014; Moravcik et al., 2017; Brown & Sandholm, 2018). In our method, after obtaining the children's instant counterfactual value of I_i, we use its behavior strategy to compute its dummy counterfactual value and update its cumulative instant regret. After that, we update its behavior strategy instantly by regret matching+ (Tammelin, 2014). Finally, we use the updated behavior strategy to update its instant counterfactual value. In previous CFR variants, the counterfactual value is only updated by the old behavior strategy rather than the latest behavior strategy.

The *average strategy* $\bar{\sigma_i}^T$ from iteration 1 to T is defined by

$$\bar{\sigma_i}^T(a|I_i)=\frac{\sum_{t=1}^T\pi_i^{\sigma^t}(I_i)\sigma_i^t(a|I_i)}{\sum_{t=1}^T\pi_i^{\sigma^t}(I_i)}, \tag{6}$$

where $\pi_i^{\sigma^t}(I_i)$ denotes the information set reach probability of I_i at t-th iteration and is used to weight the corresponding current strategy $\sigma_i^t(a|I_i)$.

Because the counterfactual value is updated instantly by the latest behavior strategy, we name our method as Instant Counterfactual Regret minimization (ICFR).

3.2. Theoretical Analysis of ICFR

In this section, we will prove the convergence for the proposed ICFR method as presented in Theorem 3. It can guarantee ICFR converge to a Nash equilibrium with a lower bound of the CFR.

Theorem 1 *(Theorem 2 in Zinkevich et al. (2007), Theorem 1 in Brown & Sandholm (2016)) In a two-player zero-sum perfect-recall IIG at iteration T, $\forall i \in N$, if the bound of average overall regret is ϵ_i, then $\bar{\sigma}^T$ is a $\epsilon_0+\epsilon_1$-Nash equilibrium.*

Before we prove the bound, we should prove Lemma 1 and Lemma 2.

Lemma 1 *$\forall \sigma'_{-i} \in \Sigma_{-i}$, $\forall I_i \in \mathcal{I}_i$, and $\forall a \in A(I_i)$,* $\sum_{a\in A(I_i)}Q_i^{t-1}(a|I_i)\hat{q}_i^{(\sigma_i^t,\sigma'_{-i})}(a|I_i)=0$

We can prove Lemma 1 in the same way as Lemma 14 in Burch (2017). Although these two Lemmas have different definitions of counterfactual value, they hold similar property. The proved Lemma 1 holds for any $\sigma'_{-i}\in\Sigma_{-i}$ and is more general than the previous proof.

Proof

$$\begin{aligned}
&\sum_{a\in A(I_i)}Q_i^{t-1}(a|I_i)\hat{q}_i^{(\sigma_i^t,\sigma'_{-i})}(a|I_i)\\
&=\sum_{a\in A(I_i)}Q_i^{t-1}(a|I_i)\left(s_i^{(\sigma_i^t,\sigma'_{-i})}(a|I_i)-\hat{s}_i^{(\sigma_i^t,\sigma'_{-i})}(I_i)\right)\\
&=\sum_{a\in A(I_i)}Q_i^{t-1}(a|I_i)\left(s_i^{(\sigma_i^t,\sigma'_{-i})}(a|I_i)-\sum_{b\in A(I_i)}s_i^{(\sigma_i^t,\sigma'_{-i})}(b|I_i)\sigma_i^t(b|I_i)\right)\\
&=\sum_{a\in A(I_i)}Q_i^{t-1}(a|I_i)s_i^{(\sigma_i^t,\sigma'_{-i})}(a|I_i)\\
&-\sum_{a\in A(I_i)}Q_i^{t-1}(a|I_i)\sum_{b\in A(I_i)}s_i^{(\sigma_i^t,\sigma'_{-i})}(b|I_i)\frac{Q_i^{t-1}(b|I_i)}{\sum_{c\in A(I_i)}Q_i^{t-1}(c|I_i)}\\
&=\sum_{a\in A(I_i)}Q_i^{t-1}(a|I_i)s_i^{(\sigma_i^t,\sigma'_{-i})}(a|I_i)\\
&-\sum_{b\in A(I_i)}s_i^{(\sigma_i^t,\sigma'_{-i})}(b|I_i)Q^{t-1}(b|I_i)\frac{\sum_{a\in A(I_i)}Q_i^{t-1}(a|I_i)}{\sum_{c\in A(I_i)}Q_i^{t-1}(c|I_i)}\\
&=0
\end{aligned} \tag{7}$$

■

Lemma 2 *Define $L=\max_{I_i,a,t}|\hat{q}^t(a|I_i)|.\forall I_i\in\mathcal{I},a\in A(I_i),t\in[1,T]$, we have $Q^T(a|I_i)\le L\sqrt{|A|T}$.*

Proof According to Lemma 1, we can prove

$$\begin{aligned}
&\sum_{a\in A(I_i)}Q_i^T(a|I_i)^2\le\sum_{a\in A(I_i)}\left(Q_i^{T-1}(a|I_i)+\hat{q}^T(a|I_i)\right)^2\\
&\le\sum_{a\in A(I_i)}\left(Q_i^{T-1}(a|I_i)^2+\hat{q}^t(a|I_i)^2\right)+2\sum_{a\in A(I_i)}Q_i^{T-1}(a|I_i)\hat{q}^T(a|I_i)\\
&\le\sum_{a\in A(I_i)}Q_i^{T-1}(a|I_i)^2+\sum_{a\in A(I_i)}\hat{q}^T(a|I_i)^2\\
&\le\sum_{t=1}^T\sum_{a\in A(I_i)}\hat{q}^t(a|I_i)^2\le\sum_{t=1}^T\sum_{a\in A(I_i)}L^2\le T|A|L^2
\end{aligned} \tag{8}$$

Therefore, we have $Q^T(a|I_i)\le L\sqrt{|A|T}$. ■

After that, we can prove the average overall regret of ICFR by Theorem 2.

Theorem 2 *Define $K=\min_{I_i\in\mathcal{I}_i}\left(s_i^{\sigma^t}(I_i)-\hat{s}_i^{\sigma^t}(I_i)\right)$. Define average overall regret of player i at iteration T by*

$$Q_i^T=\frac{1}{T}\max_{\sigma_i^*\in\Sigma_i}\sum_{t=1}^T\left(u_i^{(\sigma_i^*,\sigma_{-i}^t)}-u_i^{(\sigma_i^t,\sigma_{-i}^t)}\right), \tag{9}$$

then $Q_i^T\le|\mathcal{I}_i|(L\sqrt{|A|}/\sqrt{T}-K)$.

Proof

Define $\Delta I^x(I_i)$ as the incrementally reachable information sets after player i taking $x-th$ action from information set I_i. Note that, $\Delta I^x(I_i)$ doesn't contain the additionally visited information sets after player i taking $x-1$ actions from I_i. If these information sets are reached after taking action $a \in A(I_i)$, the set of incrementally reachable information sets is defined by $\Delta I^x(a|I_i)$. Define $I^x(I_i) = I^{x-1}(I_i) + \Delta I^x(I_i)$, where x is the depth of the subgame tree with root I_i and $D(I_i)$ is the maximum depth. $|I^0(I_i)| = 1$. Define $\sigma(I_i \to \sigma')$ as a strategy profile identical to σ except that player i always select action by σ' at I_i. Specifically, $\sigma(I_i \to a)$ refers to that player i always selects action a at I_i. According to the definition, we have $s_i^{\sigma(I_i \to a)}(I_i) = \sum_{I_i' \in \Delta I^1(a|I_i)} s_i^{\sigma}(I_i')$. Define $Q_i^T(I_i) = \frac{1}{T} \max_{\sigma' \in \Sigma_i} \sum_{t=1}^{T} \left(s_i^{\sigma^t(I_i \to \sigma')}(I_i) - s_i^{\sigma^t}(I_i) \right)$, then we have

$$
\begin{aligned}
&Q_i^T(I_i) \\
&= \frac{1}{T} \max_{\sigma' \in \Sigma_i} \max_{a \in A(I_i)} \sum_{t=1}^{T} \left(\sum_{I_i' \in \Delta I^1(a|I_i)} s_i^{\sigma^t(I_i \to \sigma')}(I_i') - s_i^{\sigma^t}(I_i) \right) \\
&= \frac{1}{T} \max_{\sigma' \in \Sigma_i} \max_{a \in A(I_i)} \sum_{t=1}^{T} \left(s_i^{\sigma^t(I_i \to a)}(I_i) - s_i^{\sigma^t}(I_i) \right. \\
&+ \left. \sum_{I_i' \in \Delta I^1(a|I_i)} \left(s_i^{\sigma^t(I_i \to \sigma')}(I_i') - s_i^{\sigma^t}(I_i') \right) \right) \\
&= \frac{1}{T} \max_{a \in A(I_i)} \sum_{t=1}^{T} \left(s_i^{\sigma^t(I_i \to a)}(I_i) - s_i^{\sigma^t}(I_i) \right) \\
&+ \frac{1}{T} \max_{\sigma' \in \Sigma_i} \max_{a \in A(I_i)} \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{I_i' \in \Delta I^1(a|I_i)} \left(s_i^{\sigma^t(I_i \to \sigma')}(I_i') - s_i^{\sigma^t}(I_i') \right) \right)
\end{aligned} \tag{10}
$$

According to Theorem 2 in (Burch et al., 2018), we have $s_i^{\sigma^t}(I_i) \geq \hat{s}_i^{\sigma^t}(I_i)$. Define $\Delta S(I_i) = s_i^{\sigma^t}(I_i) - \hat{s}_i^{\sigma^t}(I_i)$. Then we have

$$
\begin{aligned}
&Q_i^T(I_i) \\
&\leq \frac{1}{T} \max_{a \in A(I_i)} \sum_{t=1}^{T} \left(s_i^{\sigma^t(I_i \to a)}(I_i) - \hat{s}_i^{\sigma^t}(I_i) - \Delta S(I_i) \right) \\
&+ \max_{a \in A(I_i)} \sum_{I_i' \in \Delta I^1(a|I_i)} Q_i^T(I_i') \\
&\leq \frac{1}{T} \max_{a \in A(I_i)} \sum_{t=1}^{T} \left(s_i^{\sigma^t(I_i \to a)}(I_i) - \hat{s}_i^{\sigma^t}(I_i) - \Delta S(I_i) \right) \\
&+ \sum_{I_i' \in \Delta I^1(I_i)} Q_i^T(I_i')
\end{aligned} \tag{11}
$$

It is clear that equation 11 provides a recursive definition between $Q_i^T(I_i)$ and its children's $Q_i^T(I_i')$. We can derive that

$$
Q_i^T(I_i) \leq \frac{1}{T} \sum_{I_i' \in I_i^{D(I_i)}(I_i)} \max_{a \in A(I_i)} \sum_{t=1}^{T} \left(s_i^{\sigma^t(I_i \to a)}(I_i) - \hat{s}_i^{\sigma^t}(I_i) - K \right) \tag{12}
$$

According to Lemma 2, we have

$$
Q_i^T \leq \frac{1}{T} |\mathcal{I}_i| (L\sqrt{|A|T} - TK) \leq |\mathcal{I}_i| (L\sqrt{|A|}/\sqrt{T} - K) \tag{13}
$$

■

Theorem 3 *In a two-player zero-sum perfect-recall game at iteration T, ICFR approaches a $|\mathcal{I}|(L\sqrt{|A|}/\sqrt{T} - K)$-Nash equilibrium.*

Proof According to Theorem 1, in a two-player zero-sum game at iteration T, if $\forall i \in N$, the bound of average overall regret is ϵ_i, then $\bar{\sigma}^T$ is a $\epsilon_0 + \epsilon_1$ equilibrium. According to Lemma 2, $Q_i^T(I_i) \leq |\mathcal{I}_i|(L\sqrt{|A|}/\sqrt{T} - K)$, and $|\mathcal{I}_0| + |\mathcal{I}_1| = |\mathcal{I}|$, therefore ICFR approaches a $|\mathcal{I}|(L\sqrt{|A|}/\sqrt{T} - K)$-Nash equilibrium. ■

3.3. Space Complexity

In this section, we give the space complexity of the proposed ICFR in Theorem 4. Note that ICFR has the same space complexity as original CFR and CFR+.

Theorem 4 *Define $I_i \sqsubseteq Z$ if $h \in I_i, h \sqsubseteq z$. When performing CFR (Zinkevich et al., 2007) with simultaneous updates and the proposed instant updates, it requires $2\sum_{i \in N, I_i \in \mathcal{I}_i} |A(I_i)| + 2\max_{i \in N} \sum_{I_i \sqsubseteq Z} |A(I_i)|$ space. Similarly, when using alternating updates, it requires $2\sum_{i \in N, I_i \in \mathcal{I}_i} |A(I_i)| + \max_{i \in N} \sum_{I_i \sqsubseteq Z} |A(I_i)|$ space.*

(Burch, 2017) proved the space complexity for CFR and CFR+, who require $3\sum_{i \in N, I_i \in \mathcal{I}_i} |A(I_i)|$ space and $2(\sum_{i \in N, I_i \in \mathcal{I}_i} |A(I_i)| + max_{i \in N} |\mathcal{I}_i|)$ space respectively according to the Theorem 5 and Theorem 10 in Burch (2017). The presented bound is tighter than those in Burch (2017).

4. Hybrid CFR Variants

4.1. Skipping Mechanism

When performing CFR, we initialize the cumulative regret by zero, therefore the behavior strategy starts from a uniform random strategy. The average of behavior strategy profiles within all iterations will converge to a Nash equilibrium. The weighted average of iterative behavior strategy in previous iterations is highly exploitable. It is quite natural to ask a question that whether the average strategy by skipping the previous iteration can approach an approximate Nash equilibrium and obtain a better performance.

Although the similar technique is used in DeepStack (Moravcik et al., 2017), they don't prove its theoretical convergence. In this section, we prove the theoretical bound of this skipping mechanism.

Theorem 5 *Suppose we weight average strategy by skipping the first T_s iterations. Define $E = \frac{Ts}{T}$, where $0 \leq Ts \leq T$. Define $K = \min_{I_i \in \mathcal{I}_i} \left(s_i^{\sigma^t}(I_i) - \hat{s}_i^{\sigma^t}(I_i) \right)$. In a two-player* *zero-sum IIG at iteration T, ICFR with skipping mechanism approaches a $\frac{|\mathcal{I}|(L\sqrt{|A|}/\sqrt{T}-K)+2LE}{1-E}$-Nash equilibrium.*

Proof Define

$$\sigma^* = \frac{1}{T}\underset{\sigma_i^* \in \Sigma_i}{\operatorname{argmax}} \sum_{t=T_s}^{T}\left(u_i^{(\sigma_i^*,\sigma_{-i}^t)} - u_i^{(\sigma_i^t,\sigma_{-i}^t)}\right), \quad (14)$$

$$Q_i^{1:T} = \frac{1}{T}\sum_{t=1}^{T}\left(u_i^{(\sigma_i^*,\sigma_{-i}^t)} - u_i^{(\sigma_i^t,\sigma_{-i}^t)}\right), \quad (15)$$

According to the Theorem 2, we have $Q_i^{1:T} \leq |\mathcal{I}_i|(L\sqrt{|A|}/\sqrt{T}-K)$. According to the definition, we have

$$Q_i^{1:T} \geq -LE + \frac{1}{T}\sum_{t=T_s+1}^{T}\left(u_i^{(\sigma_i^{**},\sigma_{-i}^t)} - u_i^{(\sigma_i^t,\sigma_{-i}^t)}\right) \quad (16)$$

where

$$\sigma_i^{**} = \underset{\sigma_i^{**} \in \Sigma_i}{\operatorname{argmax}} \sum_{t=T_s+1}^{T}\left(u_i^{(\sigma_i^{**},\sigma_{-i}^t)} - u_i^{(\sigma_i^t,\sigma_{-i}^t)}\right) \quad (17)$$

Therefore, we have $Q_i^{T_s:T} \leq \frac{|\mathcal{I}_i|(L\sqrt{|A|}/\sqrt{T}-K)+LE}{1-E}$. According to Theorem 1, ICFR with skipping mechanism approaches a $\frac{|\mathcal{I}|(L\sqrt{|A|}/\sqrt{T}-K)+2LE}{1-E}$-Nash equilibrium. ■

According to the Theorem 5, if $E \to 0$, that is, $T_s = 0$, then the bound is same with Theorem 3. Empirically, the method with skipping mechanism approaches to an approximated Nash equilibrium more efficiently.

4.2. ICFR Variants

There are many popular CFR variants, such as CFR (Zinkevich et al., 2007), CFR+ (Tammelin, 2014) and DCFR (Brown & Sandholm, 2018).

CFR+ (Tammelin, 2014) is similar to CFR but has three differences. First, CFR+ uses regret-matching+ in place of regret matching and is more efficient than CFR empirically. Second, CFR+ uses alternating updates for only one player's cumulative strategy and another one's cumulative regret in each iteration, while CFR uses simultaneously updates for both players' cumulative strategy and regret. Third, CFR+ weights each current strategy by t rather than uniform distribution. Similarly, we use regret matching+ (Tammelin, 2014) rather than regret matching (Zinkevich et al., 2007), because regret matching+ has a better performance empirically. If we use regret matching+ in instant counterfactual regret minimization, we can name it by ICFR+.

Discounted CFR (DCFR) (Brown & Sandholm, 2018) is a general version of CFR and CFR+ by discounting both cumulative regret and average strategy. In DCFR(α,β,γ), the accumulated positive regrets are discounted by $t^\alpha/(t^\alpha+1)$, the accumulated negative regrets are discounted by $t^\beta/(t^\beta+1)$, and contributions to the average strategy are discounted by $(t/(t+1))^\gamma$ in t-th iteration. α, β, γ are the parameters in DCFR. DCFR can obtain a better convergent strategy than both CFR and CFR+ in many games empirically after specifying suitable parameters although the proved bound is larger than CFR. When we apply the proposed instant updates into DCFR, we obtain IDCFR algorithm. Similarly, if we use regret matching+ technique to compute behavior strategy, we obtain IDCFR+ algorithm.

In the experiment, we will give a detailed comparison for these different methods.

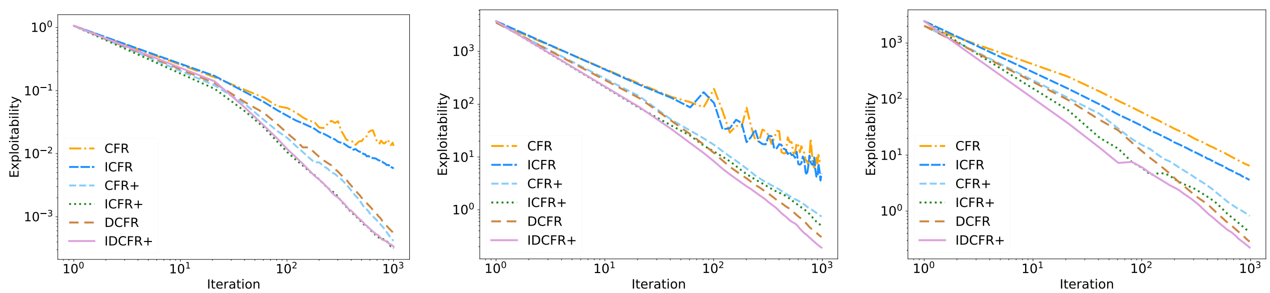
5. Experiment

We evaluated the proposed method on several different game instances: a widely-used Leduc Hold'em and five subgames of Heads-Up No-limit Texas Hold'em generated by DeepStack. The experiments cover all kinds of subgames presented in DeepStack (Moravcik et al., 2017). To reduce the randomness, we repeated each experiment for 30 times with random board and reach probability. All the experiments are evaluated by exploitability. Note that, a lower exploitability indicates better performance.

5.1. Data Sets and Game Rules

Leduc Hold'em is a two-player imperfect information game of poker and is first introduced by Southey et al. (2012). The game contains a deck of 6 cards comprising two suits of three ranks. The player may raise any amount of chips up to a maximum of that player's remaining stack. There is also no limit to the number of raises or bets in each betting round. The game has at most two rounds. In the first betting round, each player is dealt one card from a deck of 6 cards. In the second betting round, a community (or public) card is revealed from a deck of the remaining 4 cards.

Heads-up no-limit Texas hold'em (HUNL) has at most four betting rounds if neither of two players fold in advance. The four betting rounds are named by preflop, flop, turn, and river respectively. The version of HUNL we used in this paper is based on the standard of Annual Computer Poker Competition (ACPC). This version is widely used as a large data set of imperfect information game (Moravcik et al., 2017; Brown & Sandholm, 2017). Initially, both two players have 20000 chips. At the start of each hand, both players are dealt two private cards from a 52-card deck. After the preflop round, three public cards are revealed face-up on the table and the flop betting round occurs. After this round, another public card is dealt and the third betting round (called turn round) takes place. After that, the last public card is revealed, then the river round begins.



(a) Leduc Hold'em (b) River(5k) (c) River(500)

Figure 1: Convergence on (a) Leduc Hold'em (b) River(5k), and (c) River(500). It's clear that with the help of instant updates, all these methods converge faster than their original version. On Leduc Hold'em, CFR+ has a better performance than DCFR. Both the proposed ICFR+ and IDCFR+ perform better than the state-of-the-art methods. On River(5k) and River(500), IDCFR+ has a better improvement than state-of-the-art methods.

In this paper, we evaluate our methods on five subgames of HUNL produced by the DeepStack poker AI. We cover all the subgames presented in DeepStack paper (Moravcik et al., 2017). Specifically, Preflop(5k),Flop(5k), Turn(5k), River(5k) and River(500) are subgames of HUNL generated by DeepStack. The starting pot sizes for the first four subgames are 5k and the last one is 500. Note that, as the paper of DeepStack said, the terminal values of Preflop(5k) and Flop(5k) are predicted by the counterfactual value networks. The actions used to build subgames are listed in Table S3 (Moravcik et al., 2017). The exploitability is computed on each subgame. Both subgames begin at the start of the river betting round and continue to the end of the game. The start pot size of the first subgame is 500 chips and the second subgame is 5000 chips.

In this paper, we use exploitability to evaluate the performance of different methods. It's clear that the method who can obtain a lower exploitability within a specified iteration will be better.

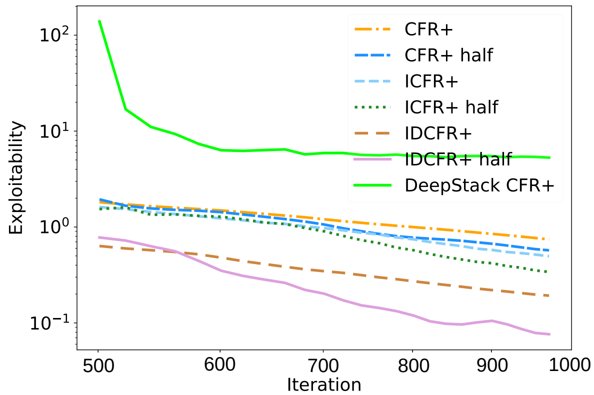
5.2. Comparison Results

When performing CFR, there are two different update methods: simultaneous method and alternating method. Empirically, the alternating method converges more efficiently than simultaneous method (Tammelin, 2014; Brown & Sandholm, 2018). In this paper, we use the alternating-updates technique on all experiments. We compared the proposed methods with the original CFR (Zinkevich et al., 2007) and the state-of-the-art methods, including CFR+ (Tammelin, 2014), DCFR (Brown & Sandholm, 2018) and hybrid CFR+ (Moravcik et al., 2017). Note that, these methods often have different performance on different game instances. Because DCFR had three different parameters, we selected these parameters by sweeping technique. Specifically, $\alpha \in [0.5,1.0,1.5,2.0,2.5]$, $\beta \in [-\infty,0,0.5,1.0,1.5,2.0,2.5]$, and $\gamma \in [1,2,3,4]$. In addition, we also applied regret matching and regret matching+ (Tammelin, 2014) into DCFR respectively. These two versions were denoted by DCFR and DCFR+ respectively.

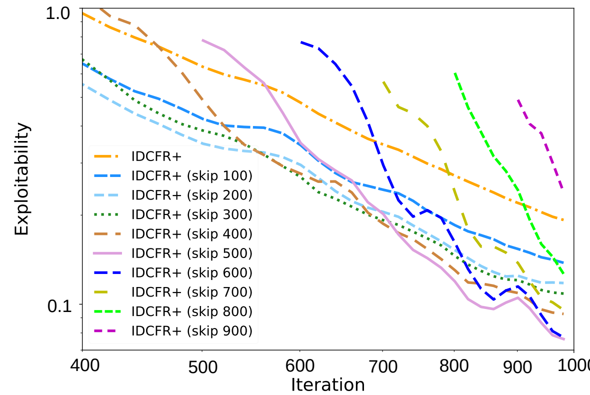
On Leduc Hold'em poker instance, Figure 1 (a) shows that CFR+ outperformed DCFR and became the strongest benchmark. With the help of the proposed instant updates, both CFR+ and IDCFR+ obtained significant improvement and converged more efficiently than the counterpart. On the subgame instances of HUNL, Figure 1 (b) and (c) shows that DCFR outperformed CFR+ and became the strongest benchmark. The proposed instant updates technique also provided a significant improvement against DCFR. When combining the proposed instant updates with the proved skipping mechanism, all of these methods converge more efficiently. Figure 2 shows that the improved IDCFR by skipping half previous iterations converges three times faster than the hybrid method used in DeepStack. In Figure 2 (b), the exploitability of IDCFR+Half between 500 and 580 iterations was larger than IDCFR+ and after 580 iterations its exploitability became lower than IDCFR+. It was reasonable because only the average strategy over a large number of iterations can approach an approximate Nash equilibrium according to the proved Theorem 5.

In practice, an exploitability of 1 mbb/g [1] is considered sufficiently converged (Michael Bowling, 2015). Thus, the performance of the presented algorithms between 100 and 1000 iterations is arguably more important than the performance beyond 10000 iterations (Moravcik et al., 2017; Brown & Sandholm, 2018). To demonstrate the performance of the proposed instant updates after long iterations, we list the performance in Table 1 for CFR and DCFR after 10k iterations. Because the performance of CFR+ and DCFR+ are much better than CFR and hybrid CFR+ empirically, we only present the performance of long iterations for CFR+ and DCFR. It's clear that instant updates technique helps both CFR+ and DCFR perform better than the counterpart.

[1] mbb/g refers to millibig blinds per game.



(a) Skipping half iterations.



(b) Skipping different iterations.

Figure 2: (a) Convergence on River(5k) for different methods by skipping half iterations. All of these methods obtain better performance with the proved skipping mechanism. (b) Convergence on River(5k) for IDCFR+ by skipping different iterations. Empirically, skipping the previous 500 or 600 iterations can obtain the best performance.

Table 1: Long-running performance after 10k iterations

Game	#infoset	#state	CFR+	ICFR+	DCFR	IDCFR+
Leduc	$1.1e4$	$6.1e4$	$2.9e{-}5{\pm}1.8e{-}7$	$\mathbf{2.3e{-}5{\pm}1.7e{-}7}$	$3.5e{-}5{\pm}4.0e{-}7$	$\mathbf{3.1e{-}5{\pm}1.6e{-}7}$
Preflop(5k)	$1.6e4$	$2.1e7$	$4.5e{-}4{\pm}4.7e{-}8$	$\mathbf{3.3e{-}4{\pm}9.6e{-}9}$	$3.0e{-}5{\pm}1.5e{-}9$	$\mathbf{9.1e{-}6{\pm}9.7e{-}9}$
Flop(5k)	$2.4e4$	$3.2e7$	$8.2e{-}3{\pm}1.2e{-}5$	$\mathbf{2.5e{-}3{\pm}5.6e{-}6}$	$1.7e{-}3{\pm}3.6e{-}6$	$\mathbf{7.3e{-}4{\pm}3.8e{-}6}$
Turn(5k)	$1.7e6$	$2.2e9$	$1.9e{-}2{\pm}3.6e{-}4$	$\mathbf{1.8e{-}2{\pm}9.5e{-}5}$	$7.7e{-}3{\pm}7.6e{-}5$	$\mathbf{7.2e{-}3{\pm}5.5e{-}5}$
River(5k)	$2.4e6$	$3.2e7$	$1.0e{-}2{\pm}1.8e{-}5$	$\mathbf{3.9e{-}3{\pm}6.0e{-}5}$	$2.2e{-}3{\pm}3.7e{-}5$	$\mathbf{1.6e{-}3{\pm}2.0e{-}5}$
River(500)	$1.6e5$	$2.1e7$	$3.4e{-}3{\pm}3.3e{-}5$	$\mathbf{2.2e{-}3{\pm}2.1e{-}5}$	$1.3e{-}3{\pm}2.5e{-}5$	$\mathbf{1.1e{-}3{\pm}3.5e{-}5}$

6. Conclusion

We have proved that counterfactual regret minimization with the proposed instant updates has a lower convergence bound. This instant updates can significantly improve the state-of-the-art method. We also have proved that weighted average strategy by skipping previous iterations approaches an approximate Nash equilibrium and helps our methods obtain a faster convergence empirically. Finally, we proved that the proposed methods have the same space complexity with CFR and the proved bound is much tighter than the proof in the previous work.

References

Brown, N. and Sandholm, T. Strategy-Based Warm Starting for Regret Minimization in Games. pp. 432–438. AAAI, 2016.

Brown, N. and Sandholm, T. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, pp. eaao1733, 2017.

Brown, N. and Sandholm, T. Solving Imperfect-Information Games via Discounted Regret Minimization. *arXiv preprint arXiv:1809.04040*, 2018.

Burch, N. Time and Space: Why Imperfect Information Games are Hard. PhD thesis, 2017.

Burch, N., Moravcik, M., and Schmid, M. Revisiting CFR+ and Alternating Updates. *arXiv preprint arXiv:1810.11542*, 2018.

Gilpin, A., Hoda, S., Pena, J., and Sandholm, T. Gradient-based algorithms for finding Nash equilibria in extensive form games. In *International Workshop on Web and Internet Economics*, pp. 57–69. Springer, 2007.

Gordon, G. J. No-regret algorithms for online convex programs. In *Advances in Neural Information Processing Systems*, pp. 489–496, 2007.

Johanson, M. Measuring the size of large no-limit poker games. *arXiv preprint arXiv:1302.7008*, 2013.

Koller, D. and Megiddo, N. The complexity of two-person zero-sum games in extensive form. *Games and economic behavior*, 4(4):528–552, 1992.

Kroer, C., Waugh, K., Kilinç-Karzan, F., and Sandholm, T. Faster first-order methods for extensive-form game solving. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pp. 817–834. ACM, 2015.

Lanctot, M., Kevin, W., Martin, Z., and Bowling, M. Monte Carlo sampling for regret minimization in extensive games. In Advances in neural information processing systems, 2009.

Li, H., Hu, K., Ge, Z., Jiang, T., Qi, Y., and Song, L. Double Neural Counterfactual Regret Minimization. *arXiv preprint arXiv:1812.10607*, 2018.

Michael Bowling, Neil Burch, M. J. O. T. Heads-Up Limit Texas Holdem is solved. *Science*, pp. 347(6218):145–149, 2015.

Moravcik, M., Martin, S., Neil, B., Viliam, L., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., and Bowling, M. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, (6337):508–513, 2017.

Osborne, M. J. and Rubinstein, A. *A course in game theory*, volume 1. MIT Press, 1994.

Schmid, M., Burch, N., Lanctot, M., Moravcik, M., Kadlec, R., and Bowling, M. Variance Reduction in Monte Carlo Counterfactual Regret Minimization (VR-MCCFR) for Extensive Form Games using Baselines. *arXiv preprint arXiv:1809.03057*, 2018.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. V. D., and et al., J. S. Mastering the game of Go with deep neural networks and tree search. *Nature*, (7587), 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

Southey, F., Bowling, M. P., Larson, B., Piccione, C., Burch, N., Billings, D., and Rayner, C. Bayes' bluff: Opponent modelling in poker. arXiv preprint arXiv:1207.1411, 2012.

Tammelin, O. Solving large imperfect information games using CFR+. arXiv preprint, 2014.

Tammelin, O., Burch, N., Johanson, M., and Bowling, M. Solving heads-up limit Texas Hold'em. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Zinkevich, M., Michael, J., Michael, B., and Piccione, C. Regret minimization in games with incomplete information. Advances in neural information processing systems, 2007.