Journal of Artificial Intelligence Research 64 (2019) 429-443 Submitted 10/18; published 2/19

Revisiting CFR$^+$ and Alternating Updates

Neil Burch BURCHN@GOOGLE.COM
Matej Moravcik MORAVCIK@GOOGLE.COM
Martin Schmid MSCHMID@GOOGLE.COM
DeepMind, 5 New Street Square,
London, EC4A 3TW

Abstract

The CFR$^+$ algorithm for solving imperfect information games is a variant of the popular CFR algorithm, with faster empirical performance on a range of problems. It was introduced with a theoretical upper bound on solution error, but subsequent work showed an error in one step of the proof. We provide updated proofs to recover the original bound.

1. Introduction

CFR$^+$ was introduced (Tammelin, 2014) as an algorithm for approximately solving imperfect information games, and was subsequently used to essentially solve the game of heads-up limit Texas Hold'em poker (Bowling, Burch, Johanson, & Tammelin, 2015). Another paper associated with the poker result gives a correctness proof for CFR$^+$, showing that approximation error approaches zero (Tammelin, Burch, Johanson, & Bowling, 2015).

CFR$^+$ is a variant of the CFR algorithm (Zinkevich, Johanson, Bowling, & Piccione, 2007), with much better empirical performance than CFR. One of the CFR$^+$ changes is switching from simultaneous updates to alternately updating a single player at a time. A crucial step in proving the correctness of both CFR and CFR$^+$ is linking regret, a hindsight measurement of performance, to exploitability, a measurement of the solution quality.

Later work pointed out a problem with the CFR$^+$ proof (Farina, Kroer, & Sandholm, 2019), noting that the CFR$^+$ proof makes reference to a folk theorem making the necessary link between regret and exploitability, but fails to satisfy the theorem's requirements due to the use of alternating updates in CFR$^+$. Farina *et al.* give an example of a sequence of updates which lead to zero regret for both players, but high exploitability.

We state a version of the folk theorem that links alternating update regret and exploitability, with an additional term in the exploitability bound relating to strategy improvement. By proving that CFR and CFR$^+$ generate improved strategies, we can give a new correctness proof for CFR$^+$, recovering the original bound on approximation error.

2. Definitions

We need a fairly large collection of definitions to get to the correctness proof. CFR and CFR$^+$ make use of the regret-matching algorithm (Hart & Mas-Colell, 2000) and regret-matching$^+$ algorithm (Tammelin, 2014), respectively, and we need to show some properties of these component algorithms. Both CFR and CFR$^+$ operate on extensive form games, a compact tree-based formalism for describing an imperfect information sequential decision making problem.

2.1 Regret-Matching and Regret-Matching$^+$

Regret-matching is an algorithm for solving the online regret minimisation problem. External regret is a hindsight measurement of how well a policy did, compared to always selecting some action. Given a set of possible actions A, a sequence of value functions $\boldsymbol{v}^t \in \mathbb{R}^{|A|}$, and sequence of policies $\boldsymbol{\sigma}^t \in \Delta^{|A|}$, the regret for an action is

$$\begin{aligned} \boldsymbol{r}^{t+1} &:= \boldsymbol{r}^t + \boldsymbol{v}^t - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t \\ \boldsymbol{r}^0 &:= \mathbf{0} \end{aligned} \tag{1}$$

An online regret minimisation algorithm specifies a policy $\boldsymbol{\sigma}^t$ based on past value functions and policies, such that $\max_a r_a^t/t \to 0$ as $t \to \infty$.

Let $x^+ := \max(x, 0)$, $\boldsymbol{x}^+ := \left[x_1^+, ..., x_n^+\right]$, and

$$\boldsymbol{\sigma}_{\text{rm}}(\boldsymbol{x}) := \begin{cases} \boldsymbol{x}^+/(\mathbf{1} \cdot \boldsymbol{x}^+) & \text{if } \exists a \text{ s.t. } x_a > 0 \\ \mathbf{1}/|A| & \text{otherwise} \end{cases} \tag{2}$$

Then for any $t \geq 0$, regret-matching uses a policy

$$\boldsymbol{\sigma}^t := \boldsymbol{\sigma}_{\text{rm}}(\boldsymbol{r}^t) \tag{3}$$

Regret-matching$^+$ is a variant of regret-matching that stores a set of non-negative regret-like values

$$\begin{aligned} \boldsymbol{q}^{t+1} &:= (\boldsymbol{q}^t + \boldsymbol{v}^t - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t)^+ \\ \boldsymbol{q}^0 &:= \mathbf{0} \end{aligned} \tag{4}$$

and uses the same regret-matching mapping from stored values to policy

$$\boldsymbol{\sigma}^t := \boldsymbol{\sigma}_{\text{rm}}(\boldsymbol{q}^t) \tag{5}$$

2.2 Extensive Form Games

An extensive form game (Von Neumann & Morgenstern, 1947) is a sequential decision-making problem where players have imperfect (asymmetric) information. The formal description of an extensive form game is given by a tuple $\langle H, P, p, \boldsymbol{\sigma}_c, u, \mathcal{I}\rangle$.

H is the set of all states h, which are a history of actions from the beginning of the game $\emptyset$. Given a history h and an action a, ha is the new state reached by taking action a at h. To denote a descendant relationship, we say $h \sqsubseteq j$ if j can be reached by some (possibly empty) sequence of actions from h, and $h \sqsubset j \iff h \sqsubseteq j, h \neq j$.

We will use $Z := \{h \in H \mid \nexists j \in H \text{ s.t. } h \sqsubset j\}$ to denote the set of terminal histories, where the game is over. We will use $Z(h) := \{z \in Z \mid h \sqsubseteq z\}$ to refer to the set of terminal histories that can be reached from some state h.

$A(h)$ gives the set of valid actions at $h \in H \setminus Z$. We assume some fixed ordering $a_1, a_2, ..., a_{|A|}$ of the actions, so we can speak about a vector of values or probabilities across actions. $a \prec b$ denotes that action a precedes b, with $a \prec b \iff a_i = a, a_j = b, i < j$.

P is the set of players, and $p : H \setminus Z \to P \bigcup \{c\}$ gives the player that is acting at state h, or the special chance player c for states where a chance event occurs according to probabilities specified by $\boldsymbol{\sigma}_c(h) \in \Delta^{|A(h)|}$. Our work is restricted to two player games, so will say $P = \{1, 2\}$.

The utility of a terminal history z for Player p is given by $u_p(z)$. We will restrict ourselves to zero-sum games, where $\sum_{p \in P} u_p(z) = 0$.

A player's imperfect information about the game state is represented by a partition $\mathcal{I}$ of states H based on player knowledge. For all information sets $I \in \mathcal{I}$ and all states $h, j \in I$ are indistinguishable to Player $p(h) = p(j)$, with the same legal actions $A(h) = A(j)$. Given this equality, we can reasonably talk about $p(I) := p(h)$ and $A(I) := A(h)$ for any $h \in I$. For any h, we will use $I(h) := I \in \mathcal{I}$ such that $h \in I$ to refer to the information set containing h. It is convenient to group information sets by the acting player, so we will use $\mathcal{I}_p := \{I \in \mathcal{I} \mid p(I) = p\}$ to refer to Player p's information sets.

We will also restrict ourselves to extensive form games where players have perfect recall. Informally, Player p has perfect recall if they do not forget anything they once knew: for all states h, j in some information set, both h and j passed through the same sequence of Player p information sets from the beginning of the game $\emptyset$, and made the same Player p actions.

A strategy $\sigma_p : \mathcal{I}_p \to \Delta^{|A(I)|}$ for Player p gives a probability distribution $\boldsymbol{\sigma}_p(I)$ over legal actions for Player p information sets. For convenience, let $\boldsymbol{\sigma}_p(h) := \boldsymbol{\sigma}_p(I(h))$. A strategy profile $\boldsymbol{\sigma} := (\sigma_1, \sigma_2)$ is a tuple of strategies for both players. Given a profile $\boldsymbol{\sigma}$, we will use σ_{-p} to refer to the strategy of p's opponent.

Because states are sequences of actions, we frequently need to refer to various products of strategy action probabilities. Given a strategy profile $\boldsymbol{\sigma}$,

$$\pi^{\boldsymbol{\sigma}}(h) := \prod_{ia \sqsubseteq h} \sigma_{p(i)}(h)_a \tag{6}$$

refers to the probability of a game reaching state h when players sample actions according to $\boldsymbol{\sigma}$ and chance events occur according to σ_c.

$$\pi^{\boldsymbol{\sigma}}(h \mid j) := \prod_{\substack{ia \sqsubseteq h \\ j \sqsubseteq i}} \sigma_{p(i)}(h)_a \tag{7}$$

refers to the probability of a game reaching h given that j was reached.

$$\begin{aligned} \pi_p^{\boldsymbol{\sigma}}(h) &:= \prod_{\substack{ia \sqsubseteq h \\ p(h) = p}} \sigma_{p(i)}(h)_a \\ \pi_{-p}^{\boldsymbol{\sigma}}(h) &:= \prod_{\substack{ia \sqsubseteq h \\ p(h) \neq p}} \sigma_{p(i)}(h)_a \end{aligned} \tag{8}$$

refer to probabilities of Player p or all actors but p making the actions to reach h, given that p's opponent and chance made the actions in h. Note that there is a slight difference in the meaning of the label $_{-p}$ here, with $\pi_{-p}^{\boldsymbol{\sigma}}$ considering actions by both Player p's opponent and chance, whereas σ_{-p} refers to the strategy of p's opponent.

$$\pi_p^{\boldsymbol{\sigma}}(h \mid j) := \prod_{\substack{ia \sqsubseteq h \\ j \sqsubseteq i \\ p(h)=p}} \sigma_{p(i)}(h)_a \tag{9}$$

refers to the probability of Player p making the actions to reach h, given j was reached and p's opponent and chance make the actions to reach h. There are a few useful relationships:

$$\begin{aligned} &\pi^{\boldsymbol{\sigma}}(h) = \pi_p^{\boldsymbol{\sigma}}(h)\pi_{-p}^{\boldsymbol{\sigma}}(h) \\ &\forall j \sqsubseteq h,\ \pi^{\boldsymbol{\sigma}}(h) = \pi^{\boldsymbol{\sigma}}(j)\pi^{\boldsymbol{\sigma}}(h \mid j) \end{aligned} \tag{10}$$

The expected utility of a strategy profile $\boldsymbol{\sigma}$ is

$$u_p^{\boldsymbol{\sigma}} := \sum_{z \in Z} \pi^{\boldsymbol{\sigma}}(z) u_p(z) \tag{11}$$

The counterfactual value of a history or information set are defined as

$$\begin{aligned} v_p^{\boldsymbol{\sigma}}(h) &:= \sum_{z \in Z(h)} \pi_{-p}^{\boldsymbol{\sigma}}(z)\pi_p^{\boldsymbol{\sigma}}(z \mid h) u_p(z) \\ \boldsymbol{v}^{\boldsymbol{\sigma}}(I) &:= \sum_{h \in I} (v_{p(h)}^{\boldsymbol{\sigma}}(ha_1), ..., v_{p(h)}^{\boldsymbol{\sigma}}(ha_{|A(I)|})) \end{aligned} \tag{12}$$

For later convenience, we will assume that for each player there exists an information set $I_p^{\emptyset}$ at the beginning of the game, containing a single state with a single action, leading to the rest of the game. This lets us say that $u_p^{\boldsymbol{\sigma}} = v^{\boldsymbol{\sigma}}(I_p^{\emptyset})_{a_0}$.

Given a sequence $\sigma_p^0, ..., \sigma_p^t$ of strategies, we denote the average strategy from a to b as

$$\bar{\sigma}_p^{[a,b]} := \sum_{i=a}^{b} \frac{\sigma_p^i}{b-a+1} \tag{13}$$

Given a sequence $\boldsymbol{\sigma}^0, ..., \boldsymbol{\sigma}^{t-1}$ of strategy profiles, we denote the average Player p regret as

$$\begin{aligned} r_p^t &:= \max_{\sigma_p^*} \sum_{i=0}^{t-1} (u_p^{(\sigma_p^*, \sigma_{-p}^i)} - u_p^{\boldsymbol{\sigma}^i})/t \\ &= \max_{\sigma_p^*} u_p^{(\sigma_p^*, \bar{\sigma}_{-p}^{[0,t-1]})} - \sum_{i=0}^{t-1} u_p^{\boldsymbol{\sigma}^i}/t \end{aligned} \tag{14}$$

The exploitability of a strategy profile $\boldsymbol{\sigma}$ is a measurement of how much expected utility each player could gain by switching their strategy:

$$\begin{aligned} \text{expl}(\boldsymbol{\sigma}) &:= \max_{\sigma_1^*} u_1^{(\sigma_1^*, \sigma_2)} - u_1^{\boldsymbol{\sigma}} + \max_{\sigma_2^*} u_2^{(\sigma_1, \sigma_2^*)} - u_2^{\boldsymbol{\sigma}} \\ &= \max_{\sigma_1^*} u_1^{(\sigma_1^*, \sigma_2)} + \max_{\sigma_2^*} u_2^{(\sigma_1, \sigma_2^*)} \qquad \text{by zero-sum} \end{aligned} \tag{15}$$

Achieving zero exploitability – a Nash equilibrium (Nash, 1950) – is possible. In two player, zero-sum games, finding a strategy with low exploitability is a reasonable goal for good play.

2.3 CFR and CFR$^+$

CFR and its variant CFR$^+$ are both algorithms for finding an extensive form game strategy with low exploitability. They are all iterative self-play algorithms that track the average of a current strategy that is based on many loosely coupled regret minimisation problems.

CFR and CFR$^+$ track regret-matching values $\boldsymbol{r}^t(I)$ or regret-matching$^+$ values $\boldsymbol{q}^t(I)$ respectively, for all $I \in \mathcal{I}$. At time t, CFR and CFR$^+$ use strategy profile $\boldsymbol{\sigma}^t(I) := \boldsymbol{\sigma}_{\mathrm{rm}}(\boldsymbol{r}^t(I))$ and $\boldsymbol{\sigma}^t(I) := \boldsymbol{\sigma}_{\mathrm{rm}}(\boldsymbol{q}^t(I))$, respectively. When doing alternating updates, with the first update done by Player 1, the values used for updating regrets are

$$\boldsymbol{v}^t(I) := \begin{cases} \boldsymbol{v}^{\boldsymbol{\sigma}^t}(I) & \text{if } p(I) = 1 \\ \boldsymbol{v}^{(\sigma_1^{t+1}, \sigma_2^t)}(I) & \text{if } p(I) = 2 \end{cases} \tag{16}$$

and the output of CFR is the profile of average strategies $(\bar{\sigma}_1^{[1,t]}, \bar{\sigma}_2^{[0,t-1]})$, while the output of CFR$^+$ is the profile of weighted average strategies $(\frac{2}{t^2+t}\sum_{i=1}^{t} i\sigma_1^i, \frac{2}{t^2+t}\sum_{i=0}^{t-1}(i+1)\sigma_2^i)$.

3. Theoretical Results

The CFR$^+$ proof of correctness (Tammelin et al., 2015) references a folk theorem that links regret and exploitability. Farina *et al.* show that the folk theorem only applies to simultaneous updates, not alternating updates, giving an example of a sequence of alternating updates with no regret but constant exploitability (Farina et al., 2019). Their observation is reproduced below using the definitions from this work.

Observation 1 *Let $P = \{X, Y\}$, $A = \{0, 1\}$, and $Z = \{00, 01, 10, 11\}$. A game consists of each player selecting one action. Let $u_X(11) = 1$, and $u_X(z) = 0$ for all $z \neq 11$. Consider the sequence of strategies $\sigma_X^t = \sigma_Y^t = t \bmod 2$, with Player X regrets computed using $\boldsymbol{v}^{(\sigma_X^t, \sigma_Y^t)}$ and Player Y regrets computed using $\boldsymbol{v}^{(\sigma_X^{t+1}, \sigma_Y^t)}$. Then at any time $2T$ the accumulated regret for both players is 0 and the average strategy is $\bar{\sigma}_X^{[1,2T]} = \bar{\sigma}_Y^{[0,2T-1]} = \mathbf{0.5}$, with exploitability $\mathrm{expl}(\bar{\sigma}_X^{[1,2T]}, \bar{\sigma}_Y^{[0,2T-1]}) = 0.5$. So both players have 0 regret, but the exploitability does not approach 0.*

As a first step in correcting the CFR$^+$ proof, we introduce an analogue of the folk theorem, linking alternating update regret and exploitability.

Theorem 1 *Let $\boldsymbol{\sigma}^t$ be the strategy profile at some time t, and r_p^t be the regrets computed using alternating updates so that Player 1 regrets are updated using $\boldsymbol{v}^{(\sigma_1^t, \sigma_2^t)}$ and Player 2 regrets are updated using $\boldsymbol{v}^{(\sigma_1^{t+1}, \sigma_2^t)}$. If the regrets are bounded by $r_p^t \leq \epsilon_p$, then the exploitability of $(\bar{\sigma}_1^{[1,t]}, \bar{\sigma}_2^{[0,t-1]})$ is bounded by $\epsilon_1 + \epsilon_2 - \frac{1}{t}\sum_{i=0}^{t-1}(u_1^{(\sigma_1^{i+1}, \sigma_2^i)} - u_1^{(\sigma_1^i, \sigma_2^i)})$.*

Proof. Consider the sum of regrets for both players, $r_1^t + r_2^t$

$$= \max_{\sigma_1^*} u_1^{(\sigma_1^*, \bar{\sigma}_2^{[0,t-1]})} - \frac{1}{t}\sum_{i=0}^{t-1} u_1^{(\sigma_1^i, \sigma_2^i)} + \max_{\sigma_2^*} u_2^{(\bar{\sigma}_1^{[1,t]}, \sigma_2^*)} - \frac{1}{t}\sum_{i=0}^{t-1} u_2^{(\sigma_1^{i+1}, \sigma_2^i)} \qquad \text{by Eq. 14}$$

$$= \mathrm{expl}\,(\bar{\sigma}_1^{[1,t]}, \bar{\sigma}_2^{[0,t-1]}) - \frac{1}{t}\sum_{i=0}^{t-1}(u_1^{(\sigma_1^i, \sigma_2^i)} + u_2^{(\sigma_1^{i+1}, \sigma_2^i)}) \qquad \text{by Eq. 15}$$

Given $r_p^t \leq \epsilon_p$ for all players p, we have $\text{expl}\,(\bar{\sigma}_1^{[1,t]}, \bar{\sigma}_2^{[0,t-1]})$

$$
\begin{aligned}
&\leq \epsilon_1 + \epsilon_2 + \frac{1}{t}\sum_{i=0}^{t-1}(u_1^{(\sigma_1^i,\sigma_2^i)} + u_2^{(\sigma_1^{i+1},\sigma_2^i)}) \\
&= \epsilon_1 + \epsilon_2 + \frac{1}{t}\sum_{i=0}^{t-1}(u_1^{(\sigma_1^i,\sigma_2^i)} - u_1^{(\sigma_1^{i+1},\sigma_2^i)}) && \text{by zero-sum} \\
&= \epsilon_1 + \epsilon_2 - \frac{1}{t}\sum_{i=0}^{t-1}(u_1^{(\sigma_1^{i+1},\sigma_2^i)} - u_1^{(\sigma_1^i,\sigma_2^i)})
\end{aligned}
$$

■

The gap between regret and exploitability in Observation 1 is now apparent as a trailing sum in Theorem 1. Each term in the sum measures the improvement in expected utility for Player 1 from time t to time $t+1$. Motivated by this sum, we show that regret-matching, CFR, and their $^+$ variants generate new policies which are not worse than the current policy. Using these constraints, we construct an updated correctness proof for CFR$^+$.

3.1 Regret-Matching and Regret-Matching$^+$ Properties

We will show that when using regret-matching or regret-matching$^+$, the expected utility $\boldsymbol{\sigma}^{t+1} \cdot \boldsymbol{v}^t$ is never less than $\boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t$. To do this, we will need to show these algorithms have a couple of other properties. We start by showing that once there is at least one positive stored regret or regret-like value, there will always be a positive stored value.

Lemma 2 *For any t, let $\boldsymbol{s}^t$ be the stored value $\boldsymbol{r}^t$ used by regret-matching or $\boldsymbol{q}^t$ used by regret-matching$^+$, and $\boldsymbol{\sigma}^t$ be the associated policy. Then for all t where $\exists a \in A$ such that $s_a^t > 0$, there $\exists b \in A$ such that $s_b^{t+1} > 0$.*

Proof. Consider any time t where $\exists a \in A$ such that $s_a^t > 0$. The policy at time t is then

$$\boldsymbol{\sigma}^t = \boldsymbol{s}^{t,+}/(\mathbf{1} \cdot \boldsymbol{s}^{t,+}) \qquad \text{by Eqs. 2, 3, 5} \tag{17}$$

Consider the stored value s_a^{t+1}. With regret-matching $s_a^{t+1} = r_a^{t+1} = r_a^t + v_a^t - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t$ by Equation 1, and with regret-matching$^+$ $s_a^{t+1} = q_a^{t+1} = (q_a^t + v_a^t - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t)^+$ by Equation 4. For both algorithms, the value of s_a^{t+1} depends on $v_a^t - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t$. There are two cases:

1. $v_a^t - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t \geq 0$

$$s_a^{t+1} > 0 \qquad \text{by Lemma assumption, Eq. 1, 4}$$

2. $v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t < 0$

$$\begin{aligned}
&\sigma^t_a(v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t) < 0 && \text{by } s^t_a > 0, \text{ Eq. 17}\\
&0 < \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t - \sigma^t_a(v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t) &&\\
&0 < \sum_{b\in A}\left(\sigma^t_b(v^t_b - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t)\right) - \sigma^t_a(v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t) && \text{by } \textstyle\sum_{b\in A}\sigma^t_b = 1\\
&0 < \sum_{b\neq a}\left(\sigma^t_b(v^t_b - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t)\right) &&\\
&\exists b \text{ s.t. } \sigma^t_b > 0,\ v^t_b - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t > 0 && \text{by } \sigma^t_{a'} \geq 0 \text{ for all } a' \in A\\
&s^t_b > 0,\ v^t_b - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t > 0 && \text{by Eq. 17}\\
&s^{t+1}_b > 0 && \text{by Eqs. 1, 4}
\end{aligned}$$

In both cases, $\exists b$ such that $s^{t+1}_b > 0$. ■

There is a corollary to Lemma 2, that regret-matching and regret-matching$^+$ never switch back to playing the default uniform random policy once they switch away from it.

Corollary 3 *When using regret-matching or regret-matching$^+$, if there exists a time t such that $\sigma^t = \boldsymbol{s}^{t,+}/(\mathbf{1} \cdot \boldsymbol{s}^{t,+})$ where $\boldsymbol{s}^t$ are the stored regrets $\boldsymbol{r}^t$ or regret-like values $\boldsymbol{q}^t$ at time t, then $\sigma^{t'} = \boldsymbol{s}^{t',+}/(\mathbf{1} \cdot \boldsymbol{s}^{t',+})$ for all $t' \geq t$.*

Proof. Assume that at some time t, $\sigma^t = \boldsymbol{s}^{t,+}/(\mathbf{1} \cdot \boldsymbol{s}^{t,+})$. We can show by induction that $\sigma^{t'} = \boldsymbol{s}^{t',+}/(\mathbf{1} \cdot \boldsymbol{s}^{t',+})$ for all $t' \geq t$. The base case $t' = t$ of the hypothesis holds by assumption. Now, assume that $\sigma^{t'} = \boldsymbol{s}^{t',+}/(\mathbf{1} \cdot \boldsymbol{s}^{t',+})$ for some time $t' \geq t$. We have

$$\begin{aligned}
&\exists a \in A \text{ s.t. } s^{t'}_a > 0 && \text{by Eq. 2}\\
&\exists b \in A \text{ s.t. } s^{t'+1}_b > 0 && \text{by Lemma 2}\\
&\sigma^{t'+1} = \boldsymbol{s}^{t'+1,+}/(\mathbf{1} \cdot \boldsymbol{s}^{t'+1,+}) && \text{by Eq. 2}
\end{aligned}$$

Therefore, by induction the hypothesis holds for all $t' \geq t$. ■

Lemma 4 *For any t, let $\boldsymbol{s}^t$ be the stored value $\boldsymbol{r}^t$ used by regret-matching or $\boldsymbol{q}^t$ used by regret-matching$^+$, and $\boldsymbol{\sigma}^t$ be the associated policy. Then for all t and $a \in A$,*
$(s^{t+1,+}_a - s^{t,+}_a)(v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t) \geq 0$.

Proof. Consider whether $v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t$ is positive. There are two cases.

1. $v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t \leq 0$

   For regret-matching, where $s^t_a = r^t_a$, we have

$$\begin{aligned}
&r^{t+1}_a = r^t_a + v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t && \text{by Eq. 1}\\
&r^{t+1}_a \leq r^t_a &&\\
&r^{t+1,+}_a \leq r^{t,+}_a &&
\end{aligned}$$

For regret-matching$^+$, where $s^t_a = q^t_a$, we have

$$\begin{aligned}
q^{t+1,+}_a &= (q^t_a + v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t)^+ && \text{by Eq. 4} \\
q^{t+1,+}_a &= (q^{t,+}_a + v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t)^+ && \text{by Eq. 4} \\
q^{t+1,+}_a &\leq q^{t,+}_a && \text{by monotonicity of } (\cdot)^+
\end{aligned}$$

Therefore for both algorithms we have

$$\begin{aligned}
s^{t+1,+}_a - s^{t,+}_a &\leq 0 \\
(s^{t+1,+}_a - s^{t,+}_a)(v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t) &\geq 0
\end{aligned}$$

2. $v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t > 0$

$$\begin{aligned}
& s^{t+1}_a = s^t_a + v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t && \text{by Eqs. 1, 4} \\
& s^{t+1}_a > s^t_a \\
& s^{t+1,+}_a \geq s^{t,+}_a \\
& (s^{t+1,+}_a - s^{t,+}_a)(v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t) \geq 0
\end{aligned}$$

In both cases, we have $(s^{t+1,+}_a - s^{t,+}_a)(v^t_a - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t) \geq 0$. ■

Theorem 5 *If $\boldsymbol{\sigma}^0, \boldsymbol{\sigma}^1, ...$ is the sequence of regret-matching or regret-matching$^+$ policies generated from a sequence of value functions $\boldsymbol{v}^0, \boldsymbol{v}^1, ...,$ then for all t, $\boldsymbol{\sigma}^{t+1} \cdot \boldsymbol{v}^t \geq \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t$.*

Proof. Let $\boldsymbol{s}^t$ be the stored value $\boldsymbol{r}^t$ used by regret-matching or $\boldsymbol{q}^t$ used by regret-matching$^+$. Consider whether all components of $\boldsymbol{s}^t$ or $\boldsymbol{s}^{t+1}$ are negative. By Lemma 2 we know that it can not be the case that $\exists a\ s^t_a > 0$ and $\forall b\ s^{t+1}_b \leq 0$. This leaves three cases.

1. $\forall a\ s^t_a \leq 0$ and $\forall a\ s^{t+1}_a \leq 0$

$$\begin{aligned}
& \boldsymbol{\sigma}^t = \boldsymbol{\sigma}^{t+1} = \mathbf{1}/|A| && \text{by Eqs. 2, 3, 5} \\
& \boldsymbol{\sigma}^{t+1} \cdot \boldsymbol{v}^t = \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t
\end{aligned}$$

2. $\forall a\ s^t_a \leq 0$ and $\exists a\ s^{t+1}_a > 0$

$$\begin{aligned}
& \boldsymbol{\sigma}^{t+1} = \boldsymbol{s}^{t+1,+}/(\mathbf{1} \cdot \boldsymbol{s}^{t+1,+}) && \text{by Eqs. 2, 3, 5} \\
& \forall b,\ \sigma^{t+1}_b > 0 \implies s^{t+1}_b > 0 \\
& \forall b,\ \sigma^{t+1}_b > 0 \implies v^t_b > \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t && \text{by Eqs. 1, 4} \\
& \sum_{b \in A} \sigma^{t+1}_b v^t_b > \sum_b \sigma^{t+1}_b \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t && \text{by } \sigma^{t+1}_b \geq 0 \\
& \boldsymbol{\sigma}^{t+1} \cdot \boldsymbol{v}^t > \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t && \text{by } \textstyle\sum_{b \in A} \sigma^{t+1}_b = 1
\end{aligned}$$

3. $\exists a\ s^t_a > 0$ and $\exists b\ s^{t+1}_b > 0$
   Let

$$\boldsymbol{\sigma}(\boldsymbol{w}) := \boldsymbol{w}^+/(\boldsymbol{w}^+ \cdot \mathbf{1}) \tag{18}$$

Then we have

$$\boldsymbol{\sigma}^t = \boldsymbol{\sigma}(\boldsymbol{s}^t),\ \boldsymbol{\sigma}^{t+1} = \boldsymbol{\sigma}(\boldsymbol{s}^{t+1}) \qquad \text{by Eqs. 2, 3, 5} \tag{19}$$

Consider any ordering $a_1, a_2, ..., a_{|A|}$ of actions such that $b \prec a$. Let

$$w^i_j := \begin{cases} s^{t+1}_j & j \leq i \\ s^t_j & j > i \end{cases} \tag{20}$$

Note that $\forall i < a$, $w^i_a = s^t_a > 0$, and $\forall i \geq a$, $w^i_b = s^{t+1}_b > 0$, so that $\boldsymbol{\sigma}(\boldsymbol{w}^i)$ is always well-defined. We can show by induction that for all $i \geq 0$

$$\boldsymbol{\sigma}(\boldsymbol{w}^i) \cdot \boldsymbol{v}^t \geq \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t \tag{21}$$

For the base case of $i = 0$, we have

$$\begin{aligned} &\boldsymbol{w}^0 = \boldsymbol{s}^t && \text{by Eq. 20} \\ &\boldsymbol{\sigma}(\boldsymbol{w}^0) \cdot \boldsymbol{v}^t = \boldsymbol{\sigma}(\boldsymbol{s}^t) \cdot \boldsymbol{v}^t && \\ &\boldsymbol{\sigma}(\boldsymbol{w}^0) \cdot \boldsymbol{v}^t = \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t && \text{by Eq. 19} \end{aligned}$$

Now assume that Equation 21 holds for some $i \geq 0$. By construction,

$$\forall j \neq i+1,\ w^{i+1}_j = w^i_j \qquad \text{by Eq. 20} \tag{22}$$

For notational convenience, let $\Delta_w := w^{i+1,+}_{i+1} - w^{i,+}_{i+1} = s^{t+1,+}_{i+1} - s^{t,+}_{i+1}$.

$$\begin{aligned} &\boldsymbol{\sigma}(\boldsymbol{w}^{i+1}) \cdot \boldsymbol{v}^t - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t \\ &= \frac{\boldsymbol{w}^{i+1,+} \cdot \boldsymbol{v}^t}{\boldsymbol{w}^{i+1,+} \cdot \mathbf{1}} - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t && \text{by Eq. 18} \\ &= \frac{\Delta_w v^t_{i+1} + \boldsymbol{w}^{i,+} \cdot \boldsymbol{v}^t}{\Delta_w + \boldsymbol{w}^{i,+} \cdot \mathbf{1}} - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t && \text{by Eqs. 20, 22} \\ &= \frac{\Delta_w v^t_{i+1} + (\boldsymbol{w}^{i,+} \cdot \mathbf{1})\boldsymbol{\sigma}(\boldsymbol{w}^i) \cdot \boldsymbol{v}^t}{\Delta_w + \boldsymbol{w}^{i,+} \cdot \mathbf{1}} - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t && \text{by Eq. 18} \\ &\geq \frac{\Delta_w v^t_{i+1} + (\boldsymbol{w}^{i,+} \cdot \mathbf{1})\boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t}{\Delta_w + \boldsymbol{w}^{i,+} \cdot \mathbf{1}} - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t && \text{by ind. hypothesis} \\ &= \frac{\Delta_w v^t_{i+1} + (\boldsymbol{w}^{i,+} \cdot \mathbf{1})\boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t}{\Delta_w + \boldsymbol{w}^{i,+} \cdot \mathbf{1}} - \frac{\Delta_w \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t + (\boldsymbol{w}^{i,+} \cdot \mathbf{1})\boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t}{\Delta_w + \boldsymbol{w}^{i,+} \cdot \mathbf{1}} && \\ &= \frac{\Delta_w (v^t_{i+1} - \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t)}{\Delta_w + \boldsymbol{w}^{i,+} \cdot \mathbf{1}} && \\ &\geq 0 && \text{by Lemma 4} \end{aligned}$$

$\boldsymbol{\sigma}(\boldsymbol{w}^{i+1}) \cdot \boldsymbol{v}^t \geq \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t$, so by induction Equation 21 holds for all $i \geq 0$. In particular, we can now say

$$\begin{aligned} &\boldsymbol{\sigma}(\boldsymbol{w}^{|A|}) \cdot \boldsymbol{v}^t \geq \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t && \\ &\boldsymbol{\sigma}(\boldsymbol{s}^{t+1}) \cdot \boldsymbol{v}^t \geq \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t && \text{by Eq. 20} \\ &\boldsymbol{\sigma}^{t+1} \cdot \boldsymbol{v}^t \geq \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t && \text{by Eq. 19} \end{aligned}$$

In all cases, we have $\boldsymbol{\sigma}^{t+1} \cdot \boldsymbol{v}^t \geq \boldsymbol{\sigma}^t \cdot \boldsymbol{v}^t$. ■

3.2 CFR and CFR$^+$ Properties

We now show that CFR and CFR$^+$ have properties that are similar to Theorem 5. After a player updates their strategy, that player's counterfactual value does not decrease for any action at any of their information sets. Similarly, the expected value of the player's new strategy does not decrease. Finally, using the property of non-decreasing value, we give an updated proof of an exploitability bound for CFR$^+$.

Lemma 6 *Let p be the player that is about to be updated in CFR or CFR$^+$ at some time t. Let σ^t_p be the current strategy for p, and σ_o be the opponent strategy σ^t_{-p} or σ^{t+1}_{-p} used by Equation 16. Then $\forall I \in \mathcal{I}_p$ and $\forall a \in A(I)$, $v^{(\sigma^{t+1}_p,\sigma_o)}(I)_a \geq v^{(\sigma^t_p,\sigma_o)}(I)_a$.*

Proof. We will use some additional terminology. Let the terminal states reached from I by action $a \in A(I)$ be

$$Z(I,a) := \bigcup_{h \in I} Z(ha) \tag{23}$$

and for any descendant state of I, we will call the ancestor h in I

$$h^I(j) := h \in I \text{ s.t. } h \sqsubseteq j \tag{24}$$

Let $D(I,a)$ be the set of information sets which are descendants of I given action $a \in A(I)$, and let $C(I,a)$ be the set of immediate children:

$$\begin{aligned} D(I,a) &:= \{J \in \mathcal{I}_{p(I)} \mid \exists h \in I, j \in J \text{ s.t. } ha \sqsubseteq j\} \\ C(I,a) &:= D(I,a) \setminus \bigcup_{J \in C(I,a), b \in A(J)} D(J,b) \end{aligned} \tag{25}$$

Note that by perfect recall, for $J \in C(I,a)$, $\exists h \in I$ such that $ha \sqsubseteq j$ for all $j \in J$: if one state in J is reached from I by action a, all states in J are reached from I by action a. Let the distance of an information set from the end of the game be

$$d(I) := \begin{cases} \max_{a \in A(I), J \in C(I,a)} (d(J)+1) & \text{if } \exists a \text{ s.t. } C(I,a) \neq \emptyset \\ 0 & \text{if } \forall a,\ C(I,a) = \emptyset \end{cases} \tag{26}$$

Using this new terminology, we can re-write

$$\begin{aligned} v^{\boldsymbol{\sigma}}_p(I)_a &= \sum_{h \in I} \sum_{z \in Z(h)} \pi^{\boldsymbol{\sigma}}_{-p}(z) \pi^{\boldsymbol{\sigma}}_p(z \mid ha) u_p(z) && \text{by Eq. 12} \\ &= \sum_{z \in Z(I,a)} \pi^{\boldsymbol{\sigma}}_{-p}(z) \pi^{\boldsymbol{\sigma}}_p(z \mid h^I(z)a) u_p(z) && \text{by Eqs. 23, 24} \end{aligned} \tag{27}$$

We will now show that $\forall i \geq 0$

$$\forall I \in \mathcal{I}_p \text{ s.t. } d(I) \leq i,\ \forall a \in A(I),\ v^{(\sigma^{t+1},\sigma_o)}(I)_a \geq v^{(\sigma^t_p,\sigma_o)}(I)_a \tag{28}$$

For the base case $i = 0$, consider any $I \in \mathcal{I}_p$ such that $d(I) = 0$. Given these assumptions,

$$\begin{aligned} &\forall a \in A(I),\ C(I,a) = \emptyset && \text{by Eqs. 25, 26} \\ &\forall \boldsymbol{\sigma},\ \forall a \in A(I),\ \forall z \in Z(I,a),\ \pi_p^{\boldsymbol{\sigma}}(z \mid h^I(z)a) = 1 && \text{by Eq. 9} \end{aligned} \tag{29}$$

Now consider $v_p^{(\sigma_p^{t+1},\sigma_o)}(I)_a$

$$\begin{aligned} &= \sum_{z \in Z(I,a)} \pi_{-p}^{(\sigma_p^{t+1},\sigma_o)}(z)\pi_p^{(\sigma_p^{t+1},\sigma_o)}(z \mid h^I(z)a)u_p(z) && \text{by Eq. 27} \\ &= \sum_{z \in Z(I,a)} \pi_{-p}^{(\sigma_p^{t},\sigma_o)}(z)\pi_p^{(\sigma_p^{t+1},\sigma_o)}(z \mid h^I(z)a)u_p(z) && \text{by Eq. 8} \\ &= \sum_{z \in Z(I,a)} \pi_{-p}^{(\sigma_p^{t},\sigma_o)}(z)\pi_p^{(\sigma_p^{t},\sigma_o)}(z \mid h^I(z)a)u_p(z) && \text{by Eq. 29} \\ &= v^{(\sigma_p^{t},\sigma_o)}(I)_a && \text{by Eq. 27} \end{aligned}$$

Assume the inductive hypothesis, Equation 28, holds for some $i \geq 0$. If $\forall I \in \mathcal{I}_p$, $d(I) \leq i$, Equation 28 trivially holds for $i+1$. Otherwise, consider any $I \in \mathcal{I}_p$ such that $d(I) = i+1$. Let $T(I,a)$ be the (possibly empty) set of terminal histories in $Z(I,a)$ that do not pass through another information set in $\mathcal{I}_{p(I)}$.

$$T(I,a) := Z(I,a) \setminus \bigcup_{J \in C(I,a), b \in A(J)} Z(J,b) \tag{30}$$

Because we require players to have perfect recall, terminal histories which pass through different child information sets are disjoint sets.

$$Z(J,b) \cap Z(J',b') = \emptyset \iff J = J', b = b'$$

Therefore, we can construct a partition $\mathcal{P}$ of $Z(I,a)$ from these disjoint sets and the terminal histories $T(I,a)$ which do not pass through any child information set.

$$\mathcal{P} := \{Z(J,b) \mid J \in C(I,a), b \in A(J)\} \cup \{T(I,a)\} \tag{31}$$

Note that by the induction assumption, because $d(I) = i+1$

$$\begin{aligned} &\forall J \in C(I,a),\ d(J) \leq i && \text{by Eqs. 25, 26} \\ &\forall J \in C(I,a), b \in A(J),\ v^{(\sigma^{t+1},\sigma_o)}(J)_b \geq v^{(\sigma_p^{t},\sigma_o)}(J)_b \end{aligned} \tag{32}$$

Given this, we have $v^{(\sigma_p^{t+1},\sigma_o)}(I)_a$

$$
\begin{aligned}
&= \sum_{z\in Z(I,a)} \pi_{-p}^{(\sigma_p^{t+1},\sigma_o)}(z)\pi_p^{(\sigma_p^{t+1},\sigma_o)}(z \mid h^I(z)a)u_p(z) && \text{by Eq. 27}\\
&= \sum_{z\in Z(I,a)} \pi_{-p}^{(\sigma_p^{t},\sigma_o)}(z)\pi_p^{(\sigma_p^{t+1},\sigma_o)}(z \mid h^I(z)a)u_p(z) && \text{by Eq. 8}\\
&= \sum_{J\in C(I,a)}\sum_{b\in A(J)}\sum_{z\in Z(J,b)} \pi_{-p}^{(\sigma_p^{t},\sigma_o)}(z)\pi_p^{(\sigma_p^{t+1},\sigma_o)}(z \mid h^I(z)a)u_p(z)\\
&\quad + \sum_{z\in T(I,a)} \pi_{-p}^{(\sigma_p^{t},\sigma_o)}(z)\pi_p^{(\sigma_p^{t+1},\sigma_o)}(z \mid h^I(z)a)u_p(z) && \text{by Eq. 31}\\
&= \sum_{J\in C(I,a)}\sum_{b\in A(J)}\sum_{z\in Z(J,b)} \pi_{-p}^{(\sigma_p^{t},\sigma_o)}(z)\pi_p^{(\sigma_p^{t+1},\sigma_o)}(z \mid h^I(z)a)u_p(z)\\
&\quad + \sum_{z\in T(I,a)} \pi_{-p}^{(\sigma_p^{t},\sigma_o)}(z)\pi_p^{(\sigma_p^{t},\sigma_o)}(z \mid h^I(z)a)u_p(z) && \text{by Eqs. 9, 30}
\end{aligned}
\tag{33}
$$

Looking at the terms inside $\sum_J$ we have

$$
\begin{aligned}
&\sum_{b\in A(J)}\sum_{z\in Z(J,b)} \pi_{-p}^{(\sigma_p^{t},\sigma_o)}(z)\pi_p^{(\sigma_p^{t+1},\sigma_o)}(z \mid h^I(z)a)u_p(z)\\
&= \sum_{b\in A(J)}\sum_{z\in Z(J,b)} \pi_{-p}^{(\sigma_p^{t},\sigma_o)}(z)\sigma_p^{t+1}(J)_b\pi_p^{(\sigma_p^{t+1},\sigma_o)}(z \mid h^J(z)b)u_p(z) && \text{by Eqs. 9, 25}\\
&= \sum_{b\in A(J)} \sigma_p^{t+1}(J)_b v^{(\sigma_p^{t+1},\sigma_o)}(J)_b && \text{by Eq. 27}\\
&= \boldsymbol{\sigma}_p^{t+1}(J)\cdot \boldsymbol{v}^{(\sigma_p^{t+1},\sigma_o)}(J)\\
&\geq \boldsymbol{\sigma}_p^{t+1}(J)\cdot \boldsymbol{v}^{(\sigma_p^{t},\sigma_o)}(J) && \text{by Eq. 32}\\
&\geq \boldsymbol{\sigma}_p^{t}(J)\cdot \boldsymbol{v}^{(\sigma_p^{t},\sigma_o)}(J) && \text{by Theorem 5}\\
&= \sum_{b\in A(J)}\sum_{z\in Z(J,b)} \pi_{-p}^{(\sigma_p^{t},\sigma_o)}(z)\pi_p^{(\sigma_p^{t},\sigma_o)}(z \mid h^I(z)a)u_p(z)
\end{aligned}
$$

Substituting the terms back into Equation 33, we have $v^{(\sigma_p^{t+1},\sigma_o)}(I)_a$

$$
\begin{aligned}
&\geq \sum_{J\in C(I,a)}\sum_{b\in A(J)}\sum_{z\in Z(J,b)} \pi_{-p}^{(\sigma_p^{t},\sigma_o)}(z)\pi_p^{(\sigma_p^{t},\sigma_o)}(z \mid h^I(z)a)u_p(z)\\
&\quad + \sum_{z\in T(I,a)} \pi_{-p}^{(\sigma_p^{t},\sigma_o)}(z)\pi_p^{(\sigma_p^{t},\sigma_o)}(z \mid h^I(z)a)u_p(z)\\
&= \sum_{z\in Z(I,a)} \pi_{-p}^{(\sigma_p^{t},\sigma_o)}(z)\pi_p^{(\sigma_p^{t},\sigma_o)}(z \mid h^I(z)a)u_p(z) && \text{by Eq. 31}\\
&= v^{(\sigma_p^{t},\sigma_o)}(I)_a && \text{by Eq. 27}
\end{aligned}
$$

Therefore Equation 28 holds for $i+1$, and by induction holds for all i. In particular, it holds for $i = \max_{I \in \mathcal{I}_p} d(I)$, and applies to all $I \in \mathcal{I}_p$. ■

Theorem 7 *Let p be the player that is about to be updated in CFR or CFR$^+$ at some time t. Let σ_p^t be the current strategy for p, and σ_o be the opponent strategy σ_{-p}^t or σ_{-p}^{t+1} used by the values defined in Equation 16. Then $u_p^{(\sigma_p^{t+1}, \sigma_o)} \geq u_p^{(\sigma_p^t, \sigma_o)}$.*

Proof. This immediately follows from Lemma 6 and $u_p^{\boldsymbol{\sigma}} = v^{\boldsymbol{\sigma}}(I_p^{\emptyset})_{a_0}$. ■

As a corollary of Theorems 1 and 7, when using alternating updates with either CFR or CFR$^+$, the average strategy $(\bar{\sigma}_1^{[1,t]}, \bar{\sigma}_2^{[0,t-1]})$ has $\mathcal{O}(\sqrt{t})$ exploitability. From the original papers, both algorithms have an $\mathcal{O}(\sqrt{t})$ regret bound, and the trailing sum in Theorem 1 is non-negative by Theorem 7. However, this only applies to a uniform average, so we need yet another theorem to bound the exploitability of the CFR$^+$ weighted average.

Theorem 8 *Let $\boldsymbol{\sigma}^t$ be the CFR$^+$ strategy profile at some time t, using alternating updates so that Player 1 regret-like values are updated using $\boldsymbol{v}^{(\sigma_1^t, \sigma_2^t)}$ and Player 2 regrets are updated using $\boldsymbol{v}^{(\sigma_1^{t+1}, \sigma_2^t)}$. Let $l = \max_{y,z \in Z}(u_1(y) - u_2(z))$ be the bound on terminal utilities. Then the exploitability of the weighted average strategy $(\frac{2}{t^2+t}\sum_{i=1}^t i\sigma_1^i, \frac{2}{t^2+t}\sum_{i=0}^{t-1}(i+1)\sigma_2^i)$ is bounded by $2|\mathcal{I}|l\sqrt{k/t}$, where $k := \max_I |A(I)|$.*

Proof. Consider two expanded sequences S^1 and S^2 of strategy profiles where the original strategy profile $\boldsymbol{\sigma}^t$ occurs $t+1$ times

$$S^1 := \underbrace{(\sigma_1^0, \sigma_2^0),}_{\text{1 copy}} \; \underbrace{(\sigma_1^1, \sigma_2^1), (\sigma_1^1, \sigma_2^1),}_{\text{2 copies}} \; \ldots, \; \underbrace{(\sigma_1^{t-1}, \sigma_2^{t-1}), \ldots, (\sigma_1^{t-1}, \sigma_2^{t-1})}_{\text{t copies}}$$

$$S^2 := \underbrace{(\sigma_1^1, \sigma_2^0),}_{\text{1 copy}} \; \underbrace{(\sigma_1^2, \sigma_2^1), (\sigma_1^2, \sigma_2^1),}_{\text{2 copies}} \; \ldots, \; \underbrace{(\sigma_1^t, \sigma_2^{t-1}), \ldots, (\sigma_1^t, \sigma_2^{t-1})}_{\text{t copies}}$$

Then with respect to S^p, the total Player p regret for any information set I and action a is

$$r_p^{\frac{t^2+t}{2}}(I)_a \leq tl\sqrt{kt} \qquad \text{by CFR}^+ \text{ Lemma 4 (Tammelin et al., 2015)}$$

and the average Player p regret is

$$\begin{aligned} r_p^{\frac{t^2+t}{2}} &\leq \frac{2}{t^2+t}\sum_{I \in \mathcal{I}_p} \max_a r^{\frac{t^2+t}{2}}(I)_a \qquad \text{by CFR Theorem 3 (Zinkevich et al., 2007)} \\ &\leq \frac{2}{t^2+t}\sum_{I \in \mathcal{I}_p} tl\sqrt{kt} \\ &\leq 2|\mathcal{I}_p|l\sqrt{k/t} \end{aligned} \tag{34}$$

Because we have two sequences of profiles, we can not directly use Theorem 1, but we can follow the same form as that proof to get

$$
\begin{aligned}
&r_1^{\frac{t^2+t}{2}} + r_2^{\frac{t^2+t}{2}} \\
&= \max_{\sigma_1^*} \sum_{i=0}^{\frac{t^2+t}{2}-1} \left(u_1^{(\sigma_1^*, S_{i,2}^1)} - u_1^{\boldsymbol{S}_i^1} \right) \frac{2}{t^2+t} + \max_{\sigma_2^*} \sum_{i=0}^{\frac{t^2+t}{2}-1} \left(u_2^{(\sigma_2^*, S_{i,1}^2)} - u_2^{\boldsymbol{S}_i^2} \right) \frac{2}{t^2+t} \\
&= \max_{\sigma_1^*} u_1^{\left(\sigma_1^*, \bar{S^1}_2^{[0,\frac{t^2+t}{2}-1]}\right)} + \max_{\sigma_2^*} u_2^{\left(\bar{S^2}_1^{[0,\frac{t^2+t}{2}-1]}, \sigma_2^*\right)} - \sum_{i=0}^{\frac{t^2+t}{2}} \left(u_1^{\boldsymbol{S}_i^1} + u_2^{\boldsymbol{S}_i^2} \right) \frac{2}{t^t+t} \\
&= \max_{\sigma_1^*} u_1^{\left(\sigma_1^*, \frac{2}{t^2+t}\sum_{i=0}^{t-1}(i+1)\sigma_2^i\right)} + \max_{\sigma_2^*} u_2^{\left(\frac{2}{t^2+t}\sum_{i=1}^{t} i\sigma_1^i, \sigma_2^*\right)} - \sum_{i=0}^{t-1} \frac{2(i+1)}{t^t+t} \left(u_1^{(\sigma_1^i,\sigma_2^i)} - u_1^{(\sigma_1^{i+1},\sigma_2^i)} \right) \\
&= \mathrm{expl}\left(\frac{2}{t^2+t} \sum_{i=1}^{t} i\sigma_1^i, \frac{2}{t^2+t} \sum_{i=0}^{t-1} (i+1)\sigma_2^i \right) - \sum_{i=0}^{t-1} \frac{2(i+1)}{t^t+t} \left(u_1^{(\sigma_1^i,\sigma_2^i)} - u_1^{(\sigma_1^{i+1},\sigma_2^i)} \right)
\end{aligned}
$$

Given Equation 34, we have $\mathrm{expl}\left(\frac{2}{t^2+t}\sum_{i=1}^{t} i\sigma_1^i, \frac{2}{t^2+t}\sum_{i=0}^{t-1}(i+1)\sigma_2^i\right)$

$$
\begin{aligned}
&\leq 2|\mathcal{I}|_1|l\sqrt{k/t} + 2|\mathcal{I}|_2|l\sqrt{k/t} + \frac{2}{t^t+t} \sum_{i=0}^{t-1} \left(i+1 \right) \left(u_1^{(\sigma_1^i,\sigma_2^i)} - u_1^{(\sigma_1^{i+1},\sigma_2^i)} \right) \\
&\leq 2|\mathcal{I}|_1|l\sqrt{k/t} + 2|\mathcal{I}|_2|l\sqrt{k/t} && \text{by Theorem 7} \\
&= 2|\mathcal{I}|l\sqrt{k/t}
\end{aligned}
$$

■

4. Conclusions

The original CFR$^+$ convergence proof makes unsupported use of the folk theorem linking regret to exploitability. We re-make the link between regret and exploitability for alternating updates, and provide a corrected CFR$^+$ convergence proof that recovers the original exploitability bound. The proof uses a specific property of CFR and CFR$^+$, where for any single player update, both algorithms are guaranteed to never generate a new strategy which is worse than the current strategy.

With a corrected proof, we once again have a theoretical guarantee of correctness to fall back on, and can safely use CFR$^+$ with alternating updates, in search of its strong empirical performance without worrying that it might be worse than CFR.

The alternating update analogue of the folk theorem also provides some theoretical motivation for the empirically observed benefit of using alternating updates. Exploitability is now bounded by the regret minus the average improvement in expected values. While we proved that the improvement is guaranteed to be non-negative for CFR and CFR$^+$, we would generally expect non-zero improvement on average, with a corresponding reduction in the bound on exploitability.

References

Bowling, M., Burch, N., Johanson, M., & Tammelin, O. (2015). Heads-up limit hold'em poker is solved. *Science*, *347*(6218), 145–149.

Farina, G., Kroer, C., & Sandholm, T. (2019). Online Convex Optimization for Sequential Decision Processes and Extensive-Form Games. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*.

Hart, S., & Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, *68*(5), 1127–1150.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, *36*(1), 48–49.

Tammelin, O. (2014). Solving large imperfect information games using CFR+. *CoRR*, *abs/1407.5042*.

Tammelin, O., Burch, N., Johanson, M., & Bowling, M. (2015). Solving heads-up limit texas hold'em. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*.

Von Neumann, J., & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton University Press.

Zinkevich, M., Johanson, M., Bowling, M., & Piccione, C. (2007). Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pp. 905–912.