No-Regret Learning in Extensive-Form Games with Imperfect Recall

Marc Lanctot[1] LANCTOT@UALBERTA.CA
Richard Gibson[1] RGGIBSON@CS.UALBERTA.CA
Neil Burch[1] NBURCH@UALBERTA.CA
Martin Zinkevich[2] MAZ@YAHOO-INC.COM
Michael Bowling[1] BOWLING@CS.UALBERTA.CA
[1]University of Alberta, Edmonton, Alberta, Canada T6G 2E8
[2]Yahoo! Inc., Sunnyvale, CA, USA, 94089

Abstract

Counterfactual Regret Minimization (CFR) is an efficient no-regret learning algorithm for decision problems modeled as extensive games. CFR's regret bounds depend on the requirement of perfect recall: players always remember information that was revealed to them and the order in which it was revealed. In games without perfect recall, however, CFR's guarantees do not apply. In this paper, we present the first regret bound for CFR when applied to a general class of games with imperfect recall. We also show that CFR applied to any abstraction belonging to our class results in a regret bound not just for the abstract game, but for the full game as well. We verify our theory and show how imperfect recall can be used to trade a small increase in regret for a significant reduction in memory in three domains: die-roll poker, phantom tic-tac-toe, and Bluff.

1. Introduction

Many real-world problems can be modeled as a repeated decision-making task. For problems involving multiple agents, one can model the repeated task as a normal-form game. When the task incorporates sequential decisions involving imperfect information or stochastic events, an extensive game is a useful alternative. In such decision problems, a typical goal is to minimize regret: the amount of utility lost by playing a past sequence of strategies, versus playing the best, stationary strategy in hindsight.

In this paper, we consider the problem of minimizing regret in an extensive game. A common approach to achieving low regret in extensive games is the Counterfactual Regret Minimization (CFR) (Zinkevich et al., 2008) algorithm. CFR uses a regret minimizer at every decision point with an alternative notion of regret, which provably minimizes regret in the entire extensive game. However, convergence is limited to games exhibiting perfect recall: players never forget information that was revealed to them, nor the order in the which the information was revealed. For games with imperfect recall, CFR's original analysis provides no general guarantees.

Imperfect recall brings about a number of complications. In games with perfect recall, every mixed strategy (probability distribution over pure strategies) has a utility-equivalent behavioral strategy (probability distribution over actions at each decision point) (Kuhn, 1953). While certain lossless imperfect recall games share this property (Kaneko & Kline, 1995), it is not true for imperfect recall games in general (Piccione & Rubinstein, 1996). In addition, the decision problem of determining if a player can assure themself a certain payoff in an imperfect recall game is NP-complete (Koller & Megiddo, 1992). Two-player zero-sum games can be solved by constructing an appropriate linear program (Koller et al., 1994) or minimizing regret (Zinkevich et al., 2008), provided the game has perfect recall. Without perfect recall, however, the problem becomes exponential in the worst case (Koller et al., 1994).

On the other hand, imperfect recall extensive games are more versatile than perfect recall games for modelling large real-world problems. While perfect recall requires all past information to be remembered, imperfect recall allows irrelevant information to be forgotten so that the size of the game is smaller. As CFR's memory requirements are linear in the size of the game, more games become feasible through imperfect recall. Despite the complications above, CFR has empirically been shown to work well when applied to imperfect recall abstractions of Texas Hold'em poker (Waugh et al., 2009b), but there is currently no theory

Appearing in *Proceedings of the 29^{th} International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

to suggest why this is so.

This paper presents theoretical groundings for applying CFR to games exhibiting imperfect recall. We define a general class of imperfect recall games and provide a bound on CFR's regret in such games. For a subset of this class, CFR minimizes average regret in the extensive game. Moreover, our results also provide regret guarantees when applying CFR to an abstract game, provided the abstract game belongs to our general class. We test our theory in three different domains: die-roll poker, phantom tic-tac-toe, and Bluff. To the best of our knowledge, this work demonstrates the first theoretically-grounded, practical use of imperfect recall in extensive games.

2. Background

An extensive-form game Γ with imperfect information (Osborne & Rubinstein, 1994) is a tuple $\langle N, A, H, Z, P, \sigma_c, u, \mathcal{I}\rangle$, where N is a finite set of **players**. A is a finite set of **actions**. H is a finite set of **histories**: a subset of the set of sequences of elements in A. A **prefix** of a history $h' \in H$ is a history $h \in H$ where h' begins with the sequence h; we denote prefix histories by $h \sqsubseteq h'$. For every $h \in H$, define $A(h) = \{a : a \in A, ha \in H\}$, the set of valid actions at history h; $P(h) \in N \cup \{c\}$ is the player to act at the history h, or *chance* if $P(h) = c$; and $H_i = \{h \mid h \in H, P(h) = i\}$. $Z \subseteq H$ is the set of **terminal histories**. A terminal history $z \in Z$ is a history where there does not exist any history $h \in H$, $h \neq z$ such that $z \sqsubseteq h$. The **utility function** $u_i : Z \to \mathbb{R}$ gives the utility to player $i \in N$ at each terminal history. If $|N| = 2$ and for all $z \in Z$, $\sum_{i \in N} u_i(z) = 0$, we say the game is **zero-sum**.

For each player $i \in N$, $\mathcal{I}_i$ is a partition of H_i with the property that $A(h) = A(h')$ whenever h and h' are in the same member of the partition. We call $\mathcal{I}_i$ the **information partition** of player i, and a set $I \in \mathcal{I}_i$ is an **information set** for player i. A player, when taking actions, cannot distinguish between two histories in the same information set. For $I \in \mathcal{I}_i$, we denote $A(I)$ as the set $A(h)$ for any $h \in I$. Define $I(h)$ to be the information set containing h. In this paper, we restrict ourselves to games where players cannot reach the same information set twice in a single game. Thus, we assume that for all $i \in N$ and $h, h' \in H_i$,

$$h \sqsubseteq h', h \neq h' \Rightarrow I(h) \neq I(h'). \tag{1}$$

Furthermore, σ_c is the fixed "strategy" of the special player *chance*. $\sigma_c(h, a)$ gives the probability that chance event a occurs at h. For all $h \in H_c$, $\sum_{a \in A(h)} \sigma_c(h, a) = 1$ and the decisions at any h are independent of the decision at any other $h' \neq h$.

Given a history h, define $X_i(h)$ to be the sequence of information set, action pairs such that $(I, a) \in X_i(h)$ if $I \in \mathcal{I}_i$ and there exists $h' \sqsubseteq h$ such that $h' \in I$ and $h'a \sqsubseteq h$. The order of the pairs in $X_i(h)$ is the order in which they occur in h. Define $X(h)$ to be the sequence of information set, action pairs belonging to all players in the order in which they occur in h, and $X_{-i}(h)$ similarly, by removing player i's information set, action pairs from $X(h)$. Also, define $X(h, h')$ to be the sequence of information set, action pairs belonging to all players that start at h and end at h' when $h \sqsubseteq h'$; if $h \not\sqsubseteq h'$, $X(h, h')$ is defined to be the empty sequence. $X_i(h, h')$ and $X_{-i}(h, h')$ are similarly defined.

Definition 1 *An extensive game has* **perfect recall** *if for every player $i \in N$, for every information set $I \in \mathcal{I}_i$, for any $h, h' \in I : X_i(h) = X_i(h')$. Otherwise, the game has* **imperfect recall**.

Intuitively, with perfect recall every player has an infallible memory: they cannot "forget" anything during a play of the game that they once knew. Hence, what a player knows at I is a composition of what the player has discovered in the past up to this point and the precise order in which information was discovered. Note that every perfect recall game satisfies equation (1), but not every imperfect recall game does.

A **(behavioral) strategy** σ_i for player i is a function such that for each history $h \in H_i$, $\sigma_i(h)$ is a probability distribution over $A(h)$. Furthermore, it is required that $\sigma_i(h) = \sigma_i(h')$ for all $h, h' \in I$, and we denote that as $\sigma_i(I)$. The set of all such strategies for player i is denoted by Σ_i. A **strategy profile** $\sigma \in \Sigma$ is a collection of strategies, one for each player, *i.e.* in a two-player game $\sigma = (\sigma_1, \sigma_2)$. By notational convention, σ_{-i} refers to the set of strategies including every strategy in σ except player i's strategy.

For any $\sigma \in \Sigma$, $i \in N \cup \{c\}$, and $h \in H$, define $\pi_i^\sigma(h) = \prod_{h'a \sqsubseteq h, P(h')=i} \sigma_i(h', a)$ to be the probability that player i plays to reach history h under σ. We can then define $\pi^\sigma(h) = \prod_{i \in N \cup \{c\}} \pi_i^\sigma(h)$ to be the probability that history h is reached under σ. Let $\pi_{-i}^\sigma(h)$ be the product of all players' contribution (including chance) except that of player i. Furthermore, let $\pi_i^\sigma(h, h')$ be the probability of player i playing to reach history h' after h, given h has occurred. Let $\pi^\sigma(h, h')$ and $\pi_{-i}^\sigma(h, h')$ be defined similarly. Finally, the expected utility of a strategy profile σ for player i is

$$u_i(\sigma) = \mathbb{E}_{z \in Z}[u_i(z)] = \sum_{z \in Z} u_i(z)\pi^\sigma(z).$$

We will say that a game $\Gamma' = \langle N, A', H, Z, P, \sigma_c, u, \mathcal{I}'\rangle$ is an **abstraction**, or an **abstract game**, of $\Gamma = \langle N, A, H, Z, P, \sigma_c, u, \mathcal{I}\rangle$ if for all $i \in N$ and $h, k \in H_i$, $A'(h) \subseteq A(h)$, and $I(h) = I(k)$ implies $I'(h) = I'(k)$. In this paper, we only consider abstractions where $A = A'$. A typical use of abstraction is to reduce the size of the game by ensuring that $|\mathcal{I}'| < |\mathcal{I}|$.

3. Example: Die-Roll Poker

We now introduce a game that we will use as a running example throughout the paper.

Die-roll poker (DRP) is a simplified two-player poker game that uses dice rather than cards. To begin, each player antes one chip to the **pot**. There are two betting rounds, where at the beginning of each round, players roll a private six-sided die. The game has imperfect information due to the players not seeing the result of the opponent's die rolls. During a betting round, a player may **fold** (forfeit the game), **call** (match the current bet), or **raise** (increase the current bet) by a fixed number of chips, with a maximum of two raises per round. In the first round, raises are worth two chips, whereas in the second round, raises are worth four chips. If both players have not folded by the end of the second round, a **showdown** occurs where the player with the largest sum of their two dice wins all of the chips in the pot.

DRP is naturally a game with perfect recall; players remember the exact sequence of bets made and the exact outcome of each die roll from both rounds. However, consider an imperfect recall version of DRP, **DRP-IR**, where at the beginning of the second round, both players forget their first die roll and only know the sum of their two dice. DRP-IR is an abstraction of DRP where any two histories are in the same abstract information set if and only if the sum of the player's private dice is the same and the sequence of betting is the same. DRP-IR has imperfect recall since histories that were distinguishable in the first round (for example, a roll of 1 and a roll of 4) are no longer distinguishable in the second round (for example, a roll of 1 followed by a roll of 5, and a roll of 4 followed by a roll of 2).

4. Counterfactual Regret Minimization

Given a sequence of strategy profiles $\sigma^1, \sigma^2, ..., \sigma^T$, the **(external) regret** for player i,

$$R_i^T = \max_{\sigma' \in \Sigma_i} \sum_{t=1}^{T} \left(u_i(\sigma', \sigma_{-i}^t) - u_i(\sigma_i^t, \sigma_{-i}^t) \right),$$

is the amount of utility player i could have gained had she played the best single strategy in hindsight for all time steps $t \in \{1, 2, ..., T\}$. An algorithm **minimizes regret**, or is a **no-regret algorithm**, for player i if the average positive regret approaches zero; *i.e.*, $\lim_{T \to \infty} R_i^{T,+}/T = 0$, where $x^+ = \max\{x, 0\}$. Having no regret is a desirable property. For example, it is well known that in a zero-sum game, if both players' average regret is bounded above by ϵ, then the average of the strategy profiles generated is a 2ϵ-Nash equilibrium.

Counterfactual Regret Minimization (CFR) is an iterative no-regret learning algorithm for extensive-form games having perfect recall. On each iteration t, CFR recursively traverses the entire game tree, computing the expected utility for player i at each information set $I \in \mathcal{I}_i$ under the current profile σ^t, assuming player i plays to reach I. This expectation is the **counterfactual value** for player i,

$$v_i(\sigma, I) = \sum_{z \in Z_I} u_i(z) \pi_{-i}^{\sigma}(z[I]) \pi^{\sigma}(z[I], z),$$

where Z_I is the set of terminal histories passing through I and $z[I]$ is the prefix of z contained in I ($z[I]$ is unique by equation (1)). For each action $a \in A(I)$, these values determine the **counterfactual regret** at iteration t, $r_i^t(I, a) = v_i(\sigma_{I \to a}^t, I) - v_i(\sigma^t, I)$, where $\sigma_{I \to a}$ is the profile σ except at I, action a is always taken. The regret $r_i^t(I, a)$ measures how much player i would rather play action a at I than play σ^t. Finally, σ^t is updated by applying regret matching (Hart & Mas-Colell, 2000) to the **immediate counterfactual regrets**, $R_i^T(I, a) = \sum_{t=1}^{T} r_i^t(I, a)$, according to

$$\sigma^{T+1}(I, a) = \frac{R_i^{T,+}(I, a)}{\sum_{b \in A(I)} R_i^{T,+}(I, b)},$$

with actions chosen uniformly at random when the denominator is zero. Regret matching is a no-regret learner that minimizes the per-information set immediate counterfactual regret (Zinkevich et al., 2008),

$$\max_{a \in A(I)} \frac{R_i^T(I, a)}{T} \leq \frac{\Delta_i \sqrt{|A(I)|}}{\sqrt{T}}, \tag{2}$$

where $\Delta_i = \max_{z, z' \in Z} u_i(z) - u_i(z')$. In games having perfect recall, minimizing the immediate counterfactual regrets at every information set in turn minimizes average regret, R_i^T/T. This is because perfect recall implies that the regret is bounded by the sum of the positive parts of the immediate counterfactual regrets (Zinkevich et al., 2008),

$$R_i^T \leq \sum_{I \in \mathcal{I}_i} \max_{a \in A(I)} R_i^{T,+}(I, a), \tag{3}$$

and thus

$$\frac{R_i^T}{T} \leq \frac{\Delta_i |\mathcal{I}_i| \sqrt{|A_i|}}{\sqrt{T}}, \tag{4}$$

where $|A_i| = \max_{I \in \mathcal{I}_i} |A(I)|$. CFR must store the immediate counterfactual regret for each information set, action pair, and thus CFR's memory requirements are $O(|\mathcal{I}_i||A_i|)$.

While equation (2) still holds in imperfect recall games, equation (3) and consequently equation (4) are not guaranteed to hold. An example game where CFR would exhibit high regret is provided in Section 7. Consequently, the regret for playing according to the CFR algorithm is unknown in general for imperfect recall games. However,

the advantage of applying CFR to DRP-IR, for example, is that this imperfect recall game contains fewer information sets than the full game, and thus less memory is required. Although DRP is a toy example and is small enough to run CFR on the full game, DRP is useful for understanding the concepts in the rest of this paper.

5. CFR with Imperfect Recall

In this section, we investigate the application of CFR to games with imperfect recall. We begin by showing that CFR minimizes regret for a class of games that we call "well-formed games." We then present a bound on the average regret for a more general class of imperfect recall games that we call "skew well-formed games."

5.1. Well-formed Games

For games $\Gamma = \langle N, A, H, Z, P, \sigma_c, u, \mathcal{I} \rangle$ and $\breve{\Gamma} = \langle N, A, H, Z, P, \sigma_c, u, \breve{\mathcal{I}} \rangle$, we say that $\breve{\Gamma}$ is a **perfect recall refinement of Γ** if $\breve{\Gamma}$ has perfect recall and Γ is an abstraction of $\breve{\Gamma}$. The information available to players in $\breve{\Gamma}$ is never forgotten, and is at least as informative as the information available to them in Γ. For example, DRP is a perfect recall refinement of DRP-IR. Every game has at least one perfect recall refinement by simply making $\breve{\Gamma}$ a perfect information game ($\breve{I} = \{h\}$ for all $\breve{I} \in \breve{\mathcal{I}}_i$). Furthermore, a perfect recall game is a perfect recall refinement of itself. For $I \in \mathcal{I}_i$, we define

$$\breve{\mathcal{P}}(I) = \{\breve{I} \mid \breve{I} \in \breve{\mathcal{I}}_i, \breve{I} \subseteq I\}$$

to be the set of all information sets in $\breve{\mathcal{I}}_i$ that are subsets of I. Note that our notion of refinement is similar to the one described by Kaneko & Kline (1995). Our definition differs in that we consider any possible refinement, whereas Kaneko & Kline consider only the coarsest such refinement.

Definition 2 *For a game Γ and a perfect recall refinement $\breve{\Gamma}$, we say that Γ is a **well-formed game with respect to $\breve{\Gamma}$** if for all $i \in N$, $I \in \mathcal{I}_i$, $\breve{I}, \breve{I}' \in \breve{\mathcal{P}}(I)$, there exists a bijection $\phi : Z_{\breve{I}} \to Z_{\breve{I}'}$ and constants $k_{\breve{I},\breve{I}'}, \ell_{\breve{I},\breve{I}'} \in [0, \infty)$ such that for all $z \in Z_{\breve{I}}$:*

(i) $u_i(z) = k_{\breve{I},\breve{I}'} u_i(\phi(z))$,

(ii) $\pi_c(z) = \ell_{\breve{I},\breve{I}'} \pi_c(\phi(z))$,

(iii) *In Γ, $X_{-i}(z) = X_{-i}(\phi(z))$, and*

(iv) *In Γ, $X_i(z[\breve{I}], z) = X_i(\phi(z)[\breve{I}'], \phi(z))$.*

*We say that Γ is a **well-formed game** if it is well-formed with respect to some perfect recall refinement.*

Recall that Z_I is the set of terminal histories containing a prefix in the information set I, and that $z[I]$ is that prefix. Intuitively, a game is well-formed if for each information set $I \in \mathcal{I}_i$, the structures around each $\breve{I}, \breve{I}' \in \breve{\mathcal{P}}(I)$ of some perfect recall refinement are isomorphic across four conditions. Conditions (i) and (ii) state that the corresponding utilities and chance frequencies at each terminal history are proportional. Condition (iii) asserts that the opponents can never distinguish the corresponding histories at any point in Γ. Finally, condition (iv) states that player i cannot distinguish between corresponding histories from $\breve{I}$ and $\breve{I}'$ until the end of the game.

Consider again DRP as a perfect recall refinement of DRP-IR. In DRP, the available actions are independent of dice outcomes, and the final utilities are only dependent on the final sum of the players' dice. Therefore, in DRP the utilities are equivalent between, for example, the terminal histories where player i rolled a 1 followed by a 5, and the terminal histories where player i rolled a 4 followed by a 2 (condition (i)). In addition, the chance probabilities of reaching each terminal history are equal (condition (ii)). Furthermore, the opponents can never distinguish between two isomorphic histories since player i's rolls are private (condition (iii)). Finally, in DRP-IR, player i never remembers the outcome of the first roll from the second round on (condition (iv)). Thus, DRP-IR is well-formed with respect to DRP, with constants $k_{\breve{I},\breve{I}'} = \ell_{\breve{I},\breve{I}'} = 1$.

Any perfect recall game is well-formed with respect to itself since $\breve{\mathcal{P}}(I) = \{I\}$, ϕ equal to the identity bijection, and $k_{\breve{I},\breve{I}'} = \ell_{\breve{I},\breve{I}'} = 1$ satisfies Definition 2. However, many imperfect recall games are also well-formed, with DRP-IR being one example. An additional example is presented in Section 6.

We now show that CFR can be applied to any well-formed game to minimize average regret. A sketch of the proof is described below, while a full proof is provided in an extended version of this paper (Lanctot et al., 2012).

Theorem 1 *If Γ is well-formed with respect to $\breve{\Gamma}$, then the average regret in $\breve{\Gamma}$ for player i of choosing strategies according to CFR in Γ is bounded by*

$$\frac{\breve{R}_i^T}{T} \leq \frac{\Delta_i K \sqrt{|A_i|}}{\sqrt{T}},$$

where $K = \sum_{I \in \mathcal{I}_i} \max_{\breve{I},\breve{I}' \in \breve{\mathcal{P}}(I)} k_{\breve{I},\breve{I}'} \ell_{\breve{I},\breve{I}'}$.

Proof sketch. One can show that conditions (i) to (iv) of Definition 2 imply that the positive regrets are proportional between any two information sets in $\breve{\Gamma}$ that are merged in the well-formed game, Γ. In other words, for all $I \in \mathcal{I}_i$, $\breve{I}, \breve{I}' \in \breve{\mathcal{P}}(I)$, and $a \in A(I)$,

$$R_i^{T,+}(\breve{I}, a) = k_{\breve{I},\breve{I}'} \ell_{\breve{I},\breve{I}'} R_i^{T,+}(\breve{I}', a).$$

Since regrets between Γ and $\breve{\Gamma}$ are additive, *i.e.*,

$$R_i^T(I,a) = \sum_{\breve{I} \in \breve{\mathcal{P}}(I)} R_i^T(\breve{I},a) \text{ for all } I \in \mathcal{I}_i,$$

the proportionality implies that minimizing regret at each $I \in \mathcal{I}_i$ minimizes regret at each $\breve{I} \in \breve{\mathcal{I}}_i$. Because $\breve{\Gamma}$ has perfect recall, applying equation (3) gives the result. ■

Since the strategy space is more expressive in $\breve{\Gamma}$ than in Γ ($\Sigma \subseteq \breve{\Sigma}$), $R_i^T \leq \breve{R}_i^T$ and thus it immediately follows that the average regret in Γ is minimized. In the case when Γ has perfect recall, because Γ is well-formed with respect to itself, Theorem 1 with $K = |\mathcal{I}_i|$ is a direct generalization of the original CFR bound in equation (4). Theorem 1 not only guarantees regret minimization for perfect recall games, but also for well-formed imperfect recall games.

5.2. Skew Well-formed Games

We now present a generalization of well-formed games to which a regret bound can still be derived.

Definition 3 *For a game Γ and a perfect recall refinement $\breve{\Gamma}$, we say that Γ is a **skew well-formed game with respect to $\breve{\Gamma}$** if for all $i \in N$, $I \in \mathcal{I}_i$, $\breve{I}, \breve{I}' \in \breve{\mathcal{P}}(I)$, there exists a bijection $\phi : Z_{\breve{I}} \to Z_{\breve{I}'}$ and constants $k_{\breve{I},\breve{I}'}, \delta_{\breve{I},\breve{I}'}, \ell_{\breve{I},\breve{I}'} \in [0,\infty)$ such that for all $z \in Z_{\breve{I}}$:*

(i) $\left|u_i(z) - k_{\breve{I},\breve{I}'} u_i(\phi(z))\right| \leq \delta_{\breve{I},\breve{I}'}$,

(ii) $\pi_c(z) = \ell_{\breve{I},\breve{I}'} \pi_c(\phi(z))$,

(iii) *In Γ, $X_{-i}(z) = X_{-i}(\phi(z))$, and*

(iv) *In Γ, $X_i(z[\breve{I}], z) = X_i(\phi(z)[\breve{I}'], \phi(z))$.*

*We say that Γ is a **skew well-formed game** if it is skew well-formed with respect to some perfect recall refinement.*

The only difference between Definitions 2 and 3 is in condition (i). While utilities must be exactly proportional in a well-formed game, in a skew well-formed game they must only be proportional up to a constant $\delta_{\breve{I},\breve{I}'}$. Note that any well-formed game is skew well-formed by setting $\delta_{\breve{I},\breve{I}'} = 0$.

For example, consider a new version of DRP called **Skew-DRP(δ)** with slightly modified payouts at the end of the game. Whenever the game reaches a showdown, player 1 receives a bonus δ times the number of chips in the pot from player 2 if player 1's second die roll was even; otherwise, no bonus is awarded. The pot is then awarded to the player with the highest dice sum as usual. Analogously, define **Skew-DRP-IR(δ)** to be the imperfect recall abstraction of Skew-DRP(δ) where in the second round, players only remember the sum of their two dice. Now, Skew-DRP-IR(δ) is not well-formed with respect to Skew-DRP(δ). To see this, note that the utilities resulting from the rolls 1,5 and the rolls 4,2 and the same sequence of betting are not exactly proportional because the second roll 5 is odd but 2 is even (utilities are off by δ times the pot size). However, Skew-DRP-IR(δ) is skew well-formed with respect to Skew-DRP(δ) with $\delta_{\breve{I},\breve{I}'} = \delta$ times the maximum pot size attainable from I.

Unfortunately, there is no guarantee that regret will be minimized by CFR in a skew well-formed game. However, we can still bound regret in a predictable manner according to the degree in which the utilities are skewed:

Theorem 2 *If Γ is skew well-formed with respect to $\breve{\Gamma}$, then the average regret in $\breve{\Gamma}$ for player i of choosing strategies according to CFR in Γ is bounded by*

$$\frac{\breve{R}_i^T}{T} \leq \frac{\Delta_i K \sqrt{|A_i|}}{\sqrt{T}} + \sum_{I \in \mathcal{I}_i} |\breve{\mathcal{P}}(I)| \delta_I,$$

where $K = \sum_{I \in \mathcal{I}_i} \max_{\breve{I},\breve{I}' \in \breve{\mathcal{P}}(I)} k_{\breve{I},\breve{I}'} \ell_{\breve{I},\breve{I}'}$ and $\delta_I = \max_{\breve{I},\breve{I}' \in \breve{\mathcal{P}}(I)} \delta_{\breve{I},\breve{I}'} \ell_{\breve{I},\breve{I}'}$.

The proof is similar to that of Theorem 1. Theorem 2 shows that as T approaches infinity, the bound on our regret approaches $\sum_{I \in \mathcal{I}_i} |\breve{\mathcal{P}}(I)| \delta_I$. Our experiments in Section 6 demonstrate that as the skew δ grows, so does our regret in Skew-DRP(δ) after a fixed number of iterations.

Remarks. Theorems 1 and 2 are, to our knowledge, the first to provide such theoretical guarantees in imperfect recall settings. However, these results are also relevant with regards to regret in the full game when CFR is applied to an abstraction. Recall that if Γ has perfect recall, then Γ is a perfect recall refinement of any (skew) well-formed abstract game. Thus, if we choose an abstraction that yields a (skew) well-formed game, then applying CFR to the abstract game achieves a bound on the average regret *in the full game*, Γ. This is true regardless of whether the abstraction exhibits perfect recall or imperfect recall. Previous counterexamples show that abstraction in general provides no guarantees in the full game (Waugh et al., 2009a). In contrast, our results show that applying CFR to an abstract game leads to bounded regret in the full game, provided we restrict ourselves to (skew) well-formed abstractions. If such an abstract game is much smaller than the full game, a significant amount of memory is saved when running CFR.

6. Empirical Evaluation

To complement our theoretical results, we apply CFR to both players simultaneously in several zero-sum imperfect recall (abstract) games, and measure the sum of the average regrets for both players in a perfect recall refinement (the full game). Along with the small DRP domain and its vari-

ants, we also consider the challenging domains of phantom tic-tac-toe and Bluff, which we now describe.

Phantom tic-tac-toe. As in regular tic-tac-toe, **phantom tic-tac-toe (PTTT)** is played on a 3-by-3 board, initially empty, where the goal is to claim three squares along the same row, column, or diagonal. However, in PTTT, players' actions are private. Each turn, a player attempts to take a square of their choice. If they fail due to the opponent having taken that square on a previous turn, the same player keeps trying to take an alternative square until they succeed. Players are not informed about how many attempts the opponent made before succeeding. The game ends immediately if there is ever a connecting line of squares belonging to the same player. The winner receives a payoff of $+1$, while the losing player receives -1. In PTTT, the total number of histories $|H| \approx 10^{10}$.

Bluff. Bluff, also known as Liar's Dice, Dudo, and Perudo, is a dice-bidding game. In our version, **Bluff(D_1,D_2)**, each die has six sides with faces 1 to 6. Each player i rolls D_i of these dice and looks at them without showing them to the opponent. Each round, players alternate by bidding on the outcome of all dice in play until one player claims that the other is bluffing (*i.e.*, claims that the bid does not hold). A bid consists of a **quantity** of dice and a **face** value. A face of 6 is considered "wild" and counts as matching any other face. For example, the bid 2x5 represents the claim that there are at least two dice with a face of 5 (or 6) among both players' dice. To place a new bid, the player must increase either the quantity or face value of the current bid; in addition, lowering the face is allowed if the quantity is increased. The player calling bluff wins the round if the opponent's last bid is incorrect, and loses otherwise. The losing player removes one of their dice from the game and a new round begins, starting with the player who won the previous round. When a player has no more dice left, they have lost the game. A utility of $+1$ is given for a win and -1 for a loss. In this paper, we restrict ourselves to the case where $D_1 = D_2 = 2$. Note that since Bluff(2,2) is a multi-round game, the expected values of Bluff(1,1) are precomputed for payoffs at the leaves of Bluff(2,1), which is then solved for leaf payoffs in the full Bluff(2,2) game. In Bluff(2,2), the total number of histories $|H| \approx 10^{10}$.

We consider several different imperfect recall abstractions for DRP, Skew-DRP(δ), PTTT, and Bluff. For the DRP games, we apply DRP-IR and Skew-DRP-IR(δ) respectively as described in Section 5. Our PTTT and Bluff experiments, however, also investigate the effects of imperfect recall beyond skew well-formed games. In the full, perfect recall version of PTTT, each player remembers the order of every failed and every successful move she makes throughout the entire game. In our first abstract game, **FOSF**, players forget the order of successive failures within the same turn. Clearly, there is an isomorphism between any two merged information sets $\breve{I}, \breve{I}' \in \breve{\mathcal{P}}(I)$ since the order of the actions does not affect the available future moves or utilities. Players still remember which turn each success and each failure occurred, and so the opponent's sequences of actions must be equal across the isomorphism. Thus, FOSF is well-formed. Our remaining PTTT abstractions, however, are not even skew well-formed. In **FOI**, players independently remember the sequence of failures and the sequence of successful actions, but not how the actions interleave. In **FOS**, players remember the order of failed actions, but not the order of successes. Finally, in **FOE**, players only know what actions they have taken and remember nothing about the order in which they were taken. FOI, FOS, and FOE are not skew well-formed because no isomorphism can preserve the order of the opponent's previous information set, action pairs (breaking condition (iii) of Definitions 2 and 3). In Bluff, we use abstractions described by Neller and Hnath (2011) that force players to forget everything except the last r bids. Similarly, these abstract games are not skew well-formed because the players forget information that the opponent could previously distinguish. The size of each DRP, PTTT, and Bluff game is given in Table 1, where we define $|\sigma| = |\{(I, a) : i \in N, I \in \mathcal{I}_i, a \in A(I)\}|$ to be the total number of information set, action pairs for all players. Note that Skew-DRP(δ) is the same size as DRP regardless of the skew, and recall that CFR requires space linear in $|\sigma|$.

Table 1. DRP, PTTT, and Bluff game sizes and properties. Here, $|\sigma|$ represents the total number of information set, action pairs for both players.

Game	Abstr.	Well-for.	$\lvert\sigma\rvert$	Savings
DRP	None	Yes	2610	—
DRP	DRP-IR	Yes	860	67.05%
PTTT	None	Yes	11695314	—
PTTT	FOSF	Yes	9347010	20.08%
PTTT	FOI	No	1147530	90.19%
PTTT	FOS	No	1484168	87.31%
PTTT	FOE	No	47818	99.59%
Bluff	None	Yes	704643030	—
Bluff	$r = 10$	No	295534218	58.06%
Bluff	$r = 8$	No	108323418	84.63%
Bluff	$r = 6$	No	22518468	96.80%
Bluff	$r = 4$	No	2329068	99.67%
Bluff	$r = 3$	No	543900	99.92%
Bluff	$r = 2$	No	97608	99.97%
Bluff	$r = 1$	No	12600	99.99%

For each game, we ran CFR on both players, meaning that each player's opponent was an identical copy of the same no-regret learner. Similar to Zinkevich *et al.* (2008), we

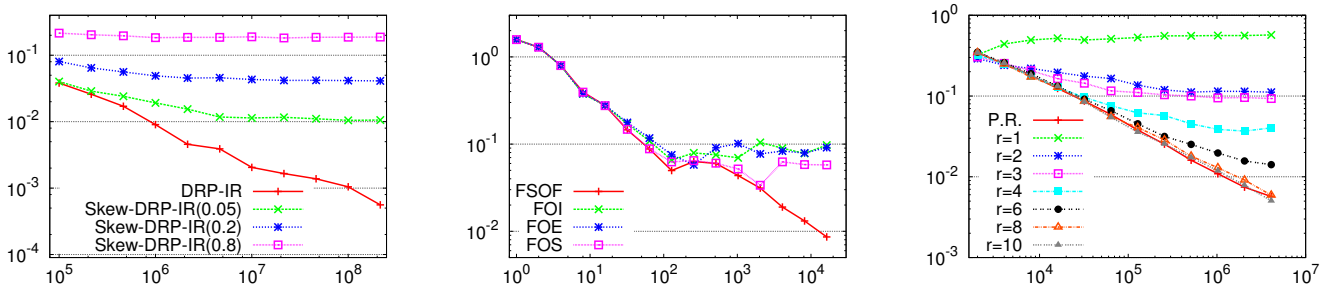

Figure 1. Sum of average positive regrets for both players as iterations increase for DRP-IR and Skew-DRP-IR(δ) (left), abstract games in PTTT (middle), and abstract games in Bluff (right). Each graph uses a log scale on both axes. The vertical axes represent the sum of average positive regrets for both players in the corresponding full unabstracted game: $(R_1^{T,+} + R_2^{T,+})/T$. The horizontal axes represent iterations T. In these graphs, P.R. is an abbreviation for perfect recall, and both FSOF and DRP-IR are well-formed.

used the chance sampling variant of CFR. The sum of the average positive regrets for each player over number of iterations is shown in Figure 1. The Skew-DRP-IR(δ) experiments show that as δ increases, so does the regret as predicted by Theorem 2, though $\sum_{I \in \mathcal{I}_i} \left|\breve{\mathcal{P}}(I)\right| \delta_I$ appears to be a very loose bound on the final regret. In PTTT, regret diverges from zero for FOI, FOS, and FOE, where FOS appears to provide slightly better strategies than FOI and FOE. While our theory cannot explain why FOS performs better, this does match our intuition that remembering information about the opponent's moves is important. For a small increase in average regret, FOS reduces the space required by 87% compared to FOSF's 20% reduction. Note that for both DRP and PTTT, running CFR on the full, perfect recall game achieves the same regret as in the well-formed abstractions (Skew-DRP-IR(0) and FSOF) and is thus not shown. In Bluff, we see that regret consistently worsens as fewer previous bids are remembered. This suggests that a result similar to Theorem 2 for skew-well-formed games may hold if condition (iii) of Definition 2 is less constrained, though the proper formulation for such a relaxation remains unclear. Nonetheless, choosing $r = 8$ saves 85% of the memory with only a very small increase in average regret after millions of iterations.

7. Discussion

Well-formed games are described by four conditions provided in Definition 2. Recall that Koller & Megiddo (1992) prove that determining a player's guaranteed payoff in an imperfect recall game is NP-complete. However, Koller & Megiddo's NP-hardness reduction creates an imperfect recall game that breaks conditions (i), (iii), and (iv) of Definition 2. In this section, we discuss the following question: For minimizing regret, how important is it to satisfy each individual condition of Definition 2?

Skew well-formed games and Theorem 2 show that one can relax condition (i) of Definition 2 and still derive a bound on the average regret. In addition, most of our PTTT and Bluff abstractions from the previous section do not satisfy condition (iii), but CFR still produces reliable results. This suggests that it may be possible to relax condition (iii) in a similar manner to the relaxation of condition (i) introduced by skew well-formed games. While we leave this question open, we now demonstrate that breaking condition (iii) can lead CFR to a dead-lock situation where one player has constant average regret.

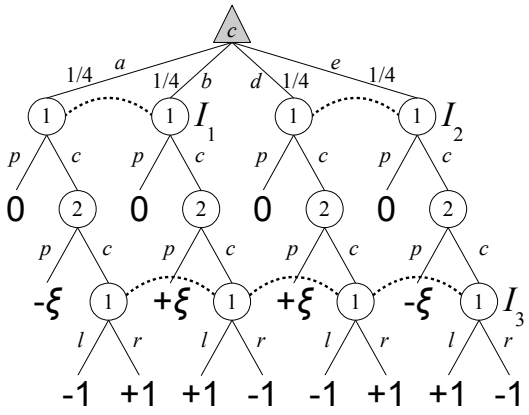

Figure 2. A zero-sum game with imperfect recall where CFR does not minimize average regret. The utilities for player 1 are given at the terminal histories, where $\xi \in (0, 1)$. Nodes connected by a bold, dashed curve are in the same information set for player 1 (player 2 has perfect information).

Let us walk through the process of applying CFR to the game in Figure 2. Note that this game satisfies all of the conditions of Definition 2, except for condition (iii). To begin, the current strategy profile σ^1 is set to be uniform random at every information set. Under this profile, when player 1 is at I_3, each of the four histories are equally likely. Thus, $v_i(\sigma^1_{(I_3 \to l)}, I_3) = v_i(\sigma^1_{(I_3 \to r)}, I_3) = v_i(\sigma^1, I_3) = 0$, and so $r_1^1(I_3, l) = r_1^1(I_3, r) = 0$. Similarly for actions p

and c at I_1 and I_2. Player 2, however, has positive immediate counterfactual regret for passing (p) at histories ac and ec (to always receive ξ utility) and for continuing (c) at bc and de (to always avoid receiving $-\xi$ utility), and has negative immediate counterfactual regret for continuing at ac and ec and for passing at bc and de. Therefore, the next profile σ^2 still has player 1 playing uniformly random everywhere, but player 2 now always passes at ac and ec, and always continues at bc and dc. On the second iteration of CFR, the positive regrets for player 1 at I_3 remain the same because the histories bcc and dcc are equally likely. Also, player 2's positive regrets remain the same at all four histories in H_2. However, player 1's expected utility for continuing at I_1 or I_2 is now negative since player 2 now passes at ac and ec, and player 1 gains positive regret for passing at both I_1 and I_2. This leads us to the next profile $\sigma^3 = \{(I_1, p) = 1, (I_2, p) = 1, (ac, p) = 1, (bc, p) = 0, (dc, p) = 0, (ec, p) = 1, (I_3, l) = 0.5\}$. One can check that running CFR for more iterations yields $\sigma^t = \sigma^3$ for all $t \geq 3$. The average regret for playing this way will be constant and hence does not approach zero because player 1 would rather play $\sigma_1' = \{(I_1, p) = 1, (I_2, p) = 0, (I_3, l) = 0\}$ and get $u_1(\sigma_1', \sigma_2^3) = (1 - \xi)/4 > u_1(\sigma^3)$ for $\xi \in (0, 1)$. A similar example can be constructed where condition (iii) holds, but chance's probabilities are not proportional (breaking condition (ii)).

Despite the problem of breaking condition (iii), condition (iv) of Definition 2 can be relaxed. Rather than enforcing player i's future information to be the same across the bijection ϕ, we only require that the corresponding subtrees be isomorphic, allowing player i to re-remember information that was previously forgotten. The details for this relaxation are in the extended version of this paper (Lanctot et al., 2012). It is not clear that this relaxation is possible in skew well-formed games, nor does it seem to provide any practical advantage.

8. Conclusion

We have provided the first set of theoretical guarantees for CFR in imperfect recall games. We defined well-formed and skew well-formed games and provided bounds on the average regret that results from applying CFR to such games. In addition, our theory shows that we can achieve low average regret in a full, perfect recall game when employing CFR on an abstract version of the game, provided the abstract game is skew well-formed (with or without imperfect recall). Our DRP experiments confirm these theoretical results, while our PTTT and Bluff experiments hint that it may be possible to still bound regret in other types of imperfect recall games. Future work will look to expand on the set of imperfect recall games to which CFR can be reliably applied. In particular, it may be possible to derive regret bounds for a new class of games where conditions (ii) and (iii) of Definition 2 are relaxed.

Acknowledgments

We would like to thank the Computer Poker Research Group at the University of Alberta for their helpful discussions. This work was supported by NSERC, Alberta Innovates – Technology Futures, and the use of computing resources provided by WestGrid and Compute Canada.

References

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

Kaneko, M. and Kline, J. J. Behavior strategies, mixed strategies and perfect recall. *International Journal of Game Theory*, 4: 127–145, 1995.

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

Koller, D., Megiddo, N., and von Stengel, B. Fast algorithms for finding randomized strategies in game trees. In *Proceedings of the 26th ACM Symposium on Theory of Computing (STOC '94)*, pp. 750–759, 1994.

Kuhn, H. W. Extensive games and the problem of information. *Contributions to the Theory of Games*, 2:193–216, 1953.

Lanctot, M., Gibson, R., Burch, N., Zinkevich, M., and Bowling, M. No-Regret Learning in Extensive-Form Games with Imperfect Recall. *ArXiv e-prints*, May 2012. Eprint 1205.0622.

Neller, T. W. and Hnath, S. Approximating optimal Dudo play with fixed-strategy iteration counterfactual regret minimization. In *Computers and Games*, 2011.

Osborne, M. J. and Rubinstein, A. *A Course in Game Theory*. MIT Press, 1994.

Piccione, M. and Rubinstein, A. On the interpretation of decision problems with imperfect recall. In *Proceedings of the 6th Conference on Theoretical Aspects of Rationality and Knowledge*, pp. 75–76. Morgan Kaufmann Publishers Inc., 1996.

Waugh, K., Schnizlein, D., Bowling, M., and Szafron, D. Abstraction pathologies in extensive games. In *The Eight International Conference on Autonomous Agents and Multiagent Systems*, pp. 781–788, 2009a.

Waugh, K., Zinkevich, M., Johanson, M., Kan, M., Schnizlein, D., and Bowling, M. A practical use of imperfect recall. In *Proceedings of SARA 2009: The Eighth Symposium on Abstraction, Reformulation and Approximation*, pp. 175–182, 2009b.

Zinkevich, M., Johanson, M., Bowling, M., and Piccione, C. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pp. 905–912, 2008.