Monte Carlo Sampling for Regret Minimization in Extensive Games

Marc Lanctot

vvichctot

variant that samples chance outcomes on each iteration [4]. They claim that the per-iteration cost reduction far exceeds the additional number of iterations required, and all of their empirical studies focus on this variant. The sampling variant and its derived bound are limited to poker-like games where chance plays a prominent role in the size of the games. This limits the practicality of CFR minimization outside of its initial application of poker or moderately sized games. An additional disadvantage of CFR is that it requires the opponent's policy to be known, which makes it unsuitable for online regret minimization in an extensive game. Online regret minimization in extensive games is possible using online convex programming techniques, such as Lagrangian Hedging [5], but these techniques can require costly optimization routines at every time step.

In this paper, we present a general framework for sampling in counterfactual regret minimization. We define a family of Monte Carlo CFR minimizing algorithms (MCCFR), that differ in how they sample the game tree on each iteration. Zinkevich's vanilla CFR and a generalization of their chance-sampled CFR are both members of this family. We then introduce two additional members of this family: *outcome-sampling*, where only a single playing of the game is sampled on each iteration; and *external-sampling*, which samples chance nodes and the opponent's actions. We show that under a reasonable sampling strategy, any member of this family minimizes overall regret, and so can be used for equilibrium computation. Additionally, external-sampling is proven to require only a constant-factor increase in iterations yet achieves an order reduction in the cost per iteration, thus resulting an asymptotic improvement in equilibrium computation time. Furthermore, since outcome-sampling does not need knowledge of the opponent's strategy beyond samples of play from the strategy, we describe how it can be used for online regret minimization. We then evaluate these algorithms empirically by using them to compute approximate equilibria in a variety of games.

2 Background

An extensive game is a general model of sequential decision-making with imperfect information. As with perfect information games (such as Chess or Checkers), extensive games consist primarily of a game tree: each non-terminal node has an associated player (possibly chance) that makes the decision at that node, and each terminal node has associated utilities for the players. Additionally, game states are partitioned into information sets where a player cannot distinguish between two states in the same information set. The players, therefore, must choose actions with the same distribution at each state in the same information set. We now define an extensive game formally, introducing the notation we use throughout the paper.

Definition 1 *[6, p. 200] a finite extensive game with imperfect information has the following components:*

- *A finite set N of* ***players****. A finite set H of sequences, the possible* ***histories*** *of actions, such that the empty sequence is in H and every prefix of a sequence in H is also in H. Define $h \sqsubseteq h'$ to mean h is a prefix of h'. $Z \subseteq H$ are the terminal histories (those which are not a prefix of any other sequences). $A(h) = \{a : ha \in H\}$ are the actions available after a non-terminal history, $h \in H \setminus Z$.*
- *A function P that assigns to each non-terminal history a member of $N \cup \{c\}$. P is the* ***player function****. $P(h)$ is the player who takes an action after the history h. If $P(h) = c$ then chance determines the action taken after history h.*
- *For each player $i \in N \cup \{c\}$ a partition $\mathcal{I}_i$ of $\{h \in H : P(h) = i\}$ with the property that $A(h) = A(h')$ whenever h and h' are in the same member of the partition. For $I_i \in \mathcal{I}_i$ we denote by $A(I_i)$ the set $A(h)$ and by $P(I_i)$ the player $P(h)$ for any $h \in I_i$. $\mathcal{I}_i$ is the* ***information partition*** *of player i; a set $I_i \in \mathcal{I}_i$ is an* ***information set*** *of player i.*
- *A function f_c that associates with every information set I where $P(I) = c$ a probability measure $f_c(\cdot|I)$ on $A(h)$ ($f_c(a|I)$ is the probability that a occurs given some $h \in I$), where each such probability measure is independent of every other such measure.*[1]

[1] Traditionally, an information partition is not specified for chance. In fact, as long as the same chance information set cannot be revisited, it has no strategic effect on the game itself. However, this extension allows us to consider using the same sampled chance outcome for an entire set of histories, which is an important part of Zinkevich and colleagues' chance-sampling CFR variant.

- *For each player $i \in N$ a utility function u_i from the terminal states Z to the reals* $\mathbf{R}$. *If $N = \{1, 2\}$ and $u_1 = -u_2$, it is a* ***zero-sum extensive game***. *Define $\Delta_{u,i} = \max_z u_i(z) - \min_z u_i(z)$ to be the range of utilities to player i.*

In this paper, we will only concern ourselves with two-player, zero-sum extensive games. Furthermore, we will assume **perfect recall**, a restriction on the information partitions such that a player can always distinguish between game states where they previously took a different action or were previously in a different information set.

2.1 Strategies and Equilibria

A **strategy of player i**, σ_i, in an extensive game is a function that assigns a distribution over $A(I_i)$ to each $I_i \in \mathcal{I}_i$. We denote Σ_i as the set of all strategies for player i. A **strategy profile**, σ, consists of a strategy for each player, $\sigma_1, \ldots, \sigma_n$. We let σ_{-i} refer to the strategies in σ excluding σ_i.

Let $\pi^\sigma(h)$ be the probability of history h occurring if all players choose actions according to σ. We can decompose $\pi^\sigma(h) = \Pi_{i \in N \cup \{c\}} \pi_i^\sigma(h)$ into each player's contribution to this probability. Here, $\pi_i^\sigma(h)$ is the contribution to this probability from player i when playing according to σ. Let $\pi_{-i}^\sigma(h)$ be the product of all players' contribution (including chance) except that of player i. For $I \subseteq H$, define $\pi^\sigma(I) = \sum_{h \in I} \pi^\sigma(h)$, as the probability of reaching a particular information set given all players play according to σ, with $\pi_i^\sigma(I)$ and $\pi_{-i}^\sigma(I)$ defined similarly. Finally, let $\pi^\sigma(h, z) = \pi^\sigma(z)/\pi^\sigma(h)$ if $h \sqsubseteq z$, and zero otherwise. Let $\pi_i^\sigma(h, z)$ and $\pi_{-i}^\sigma(h, z)$ be defined similarly. Using this notation, we can define the expected payoff for player i as $u_i(\sigma) = \sum_{h \in Z} u_i(h)\pi^\sigma(h)$.

Given a strategy profile, σ, we define a player's **best response** as a strategy that maximizes their expected payoff assuming all other players play according to σ. The **best-response value** for player i is the value of that strategy, $b_i(\sigma_{-i}) = \max_{\sigma_i' \in \Sigma_i} u_i(\sigma_i', \sigma_{-i})$. An **$\epsilon$-Nash equilibrium** is an approximation of a Nash equilibrium; it is a strategy profile σ that satisfies

$$\forall i \in N \quad u_i(\sigma) + \epsilon \geq \max_{\sigma_i' \in \Sigma_i} u_i(\sigma_i', \sigma_{-i}) \tag{1}$$

If $\epsilon = 0$ then σ is a **Nash Equilibrium**: no player has any incentive to deviate as they are all playing best responses. If a game is two-player and zero-sum, we can use **exploitability** as a metric for determining how close σ is to an equilibrium, $\epsilon_\sigma = b_1(\sigma_2) + b_2(\sigma_1)$.

2.2 Counterfactual Regret Minimization

Regret is an online learning concept that has triggered a family of powerful learning algorithms. To define this concept, first consider repeatedly playing an extensive game. Let σ_i^t be the strategy used by player i on round t. The **average overall regret** of player i at time T is:

$$R_i^T = \frac{1}{T} \max_{\sigma_i^* \in \Sigma_i} \sum_{t=1}^{T} \left(u_i(\sigma_i^*, \sigma_{-i}^t) - u_i(\sigma^t) \right) \tag{2}$$

Moreover, define $\bar{\sigma}_i^t$ to be the **average strategy** for player i from time 1 to T. In particular, for each information set $I \in \mathcal{I}_i$, for each $a \in A(I)$, define:

$$\bar{\sigma}_i^t(a|I) = \frac{\sum_{t=1}^{T} \pi_i^{\sigma^t}(I)\sigma^t(a|I)}{\sum_{t=1}^{T} \pi_i^{\sigma^t}(I)}. \tag{3}$$

There is a well-known connection between regret, average strategies, and Nash equilibria.

Theorem 1 *In a zero-sum game, if $R_{i \in \{1,2\}}^T \leq \epsilon$, then $\bar{\sigma}^T$ is a 2ϵ equilibrium.*

An algorithm for selecting σ_i^t for player i is regret minimizing if player i's average overall regret (regardless of the sequence σ_{-i}^t) goes to zero as t goes to infinity. Regret minimizing algorithms in self-play can be used as a technique for computing an approximate Nash equilibrium. Moreover, an algorithm's bounds on the average overall regret bounds the convergence rate of the approximation.

Zinkevich and colleagues [1] used the above approach in their counterfactual regret algorithm (CFR). The basic idea of CFR is that overall regret can be bounded by the sum of positive per-information-set immediate counterfactual regret. Let I be an information set of player i. Define $\sigma_{(I \to a)}$ to be a strategy profile identical to σ except that player i always chooses action a from information set I. Let Z_I be the subset of all terminal histories where a prefix of the history is in the set I; for $z \in Z_I$ let $z[I]$ be that prefix. Since we are restricting ourselves to perfect recall games $z[I]$ is unique. Define **counterfactual value** $v_i(\sigma, I)$ as,

$$v_i(\sigma, I) = \sum_{z \in Z_I} \pi^\sigma_{-i}(z[I])\pi^\sigma(z[I], z)u_i(z). \tag{4}$$

The **immediate counterfactual regret** is then $R^T_{i,\mathrm{imm}}(I) = \max_{a \in A(I)} R^T_{i,\mathrm{imm}}(I, a)$, where

$$R^T_{i,\mathrm{imm}}(I, a) = \frac{1}{T}\sum_{t=1}^{T}\Big(v_i(\sigma^t_{(I\to a)}, I) - v_i(\sigma^t, I)\Big) \tag{5}$$

Let $x^+ = \max(x, 0)$. The key insight of CFR is the following result.

Theorem 2 *[1, Theorem 3]* $R^T_i \le \sum_{I \in \mathcal{I}_i} R^{T,+}_{i,\mathrm{imm}}(I)$

Using regret-matching[2] the positive per-information set immediate counterfactual regrets can be driven to zero, thus driving average overall regret to zero. This results in an average overall regret bound [1, Theorem 4]: $R^T_i \le \Delta_{u,i}|\mathcal{I}_i|\sqrt{|A_i|}/\sqrt{T}$, where $|A_i| = \max_{h:P(h)=i}|A(h)|$. We return to this bound, tightening it further, in Section 4.

This result suggests an algorithm for computing equilibria via self-play, which we will refer to as *vanilla CFR*. The idea is to traverse the game tree computing counterfactual values using Equation 4. Given a strategy, these values define regret terms for each player for each of their information sets using Equation 5. These regret values accumulate and determine the strategies at the next iteration using the regret-matching formula. Since both players are regret minimizing, Theorem 1 applies and so computing the strategy profile $\bar{\sigma}^t$ gives us an approximate Nash Equilibrium. Since CFR only needs to store values at each information set, its space requirement is $O(|\mathcal{I}|)$. However, as previously mentioned vanilla CFR requires a complete traversal of the game tree on each iteration, which prohibits its use in many large games. Zinkevich and colleagues [4] made steps to alleviate this concern with a chance-sampled variant of CFR for poker-like games.

3 Monte Carlo CFR

The key to our approach is to avoid traversing the entire game tree on each iteration while still having the immediate counterfactual regrets be unchanged *in expectation*. In general, we want to restrict the terminal histories we consider on each iteration. Let $\mathcal{Q} = \{Q_1, \ldots, Q_r\}$ be a set of subsets of Z, such that their union spans the set Z. We will call one of these subsets a **block**. On each iteration we will sample one of these blocks and only consider the terminal histories in that block. Let $q_j > 0$ be the probability of considering block Q_j for the current iteration (where $\sum_{j=1}^r q_j = 1$).

Let $q(z) = \sum_{j:z\in Q_j} q_j$, *i.e.,* $q(z)$ is the probability of considering terminal history z on the current iteration. The **sampled counterfactual value** when updating block j is:

$$\tilde{v}_i(\sigma, I|j) = \sum_{z \in Q_j \cap Z_I} \frac{1}{q(z)} u_i(z)\pi^\sigma_{-i}(z[I])\pi^\sigma(z[I], z) \tag{6}$$

Selecting a set $\mathcal{Q}$ along with the sampling probabilities defines a complete sample-based CFR algorithm. Rather than doing full game tree traversals the algorithm samples one of these blocks, and then examines only the terminal histories in that block.

Suppose we choose $\mathcal{Q} = \{Z\}$, *i.e.,* one block containing all terminal histories and $q_1 = 1$. In this case, sampled counterfactual value is equal to counterfactual value, and we have vanilla CFR. Suppose instead we choose each block to include all terminal histories with the same sequence of chance outcomes (where the probability of a chance outcome is independent of players' actions as

[2] Regret-matching selects actions with probability proportional to their positive regret, *i.e.,* $\sigma^t_i(a|I) = R^{T,+}_{i,\mathrm{imm}}(I, a)/\sum_{a'\in A(I)} R^{T,+}_{i,\mathrm{imm}}(I, a)$. Regret-matching satisfies Blackwell's approachability criteria. [7, 8]

in poker-like games). Hence q_j is the product of the probabilities in the sampled sequence of chance outcomes (which cancels with these same probabilities in the definition of counterfactual value) and we have Zinkevich and colleagues' chance-sampled CFR.

Sampled counterfactual value was designed to match counterfactual value on expectation. We show this here, and then use this fact to prove a probabilistic bound on the algorithm's average overall regret in the next section.

Lemma 1 $E_{j \sim q_j}\left[\tilde{v}_i(\sigma, I|j)\right] = v_i(\sigma, I)$

Proof:

$$\begin{aligned} E_{j \sim q_j}\left[\tilde{v}_i(\sigma, I|j)\right] &= \sum_j q_j \tilde{v}_i(\sigma, I|j) = \sum_j \sum_{z \in Q_j \cap Z_I} \frac{q_j}{q(z)} \pi^\sigma_{-i}(z[I]) \pi^\sigma(z[I], z) u_i(z) && (7) \\ &= \sum_{z \in Z_I} \frac{\sum_{j : z \in Q_j} q_j}{q(z)} \pi^\sigma_{-i}(z[I]) \pi^\sigma(z[I], z) u_i(z) && (8) \\ &= \sum_{z \in Z_I} \pi^\sigma_{-i}(z[I]) \pi^\sigma(z[I], z) u_i(z) = v_i(\sigma, I) && (9) \end{aligned}$$

Equation 8 follows from the fact that $\mathcal{Q}$ spans Z. Equation 9 follows from the definition of $q(z)$. ■

This results in the following MCCFR algorithm. We sample a block and for each information set that contains a prefix of a terminal history in the block we compute the *sampled immediate counterfactual regrets* of each action, $\tilde{r}(I, a) = \tilde{v}_i(\sigma^t_{(I \to a)}, I) - \tilde{v}_i(\sigma^t, I)$. We accumulate these regrets, and the player's strategy on the next iteration applies the regret-matching algorithm to the accumulated regrets. We now present two specific members of this family, giving details on how the regrets can be updated efficiently.

Outcome-Sampling MCCFR. In *outcome-sampling MCCFR* we choose $\mathcal{Q}$ so that each block contains a single terminal history, *i.e.,* $\forall Q \in \mathcal{Q}, |Q| = 1$. On each iteration we sample one terminal history and only update each information set along that history. The sampling probabilities, q_j must specify a distribution over terminal histories. We will specify this distribution using a *sampling profile*, σ', so that $q(z) = \pi^{\sigma'}(z)$. Note that any choice of sampling policy will induce a particular distribution over the block probabilities $q(z)$. As long as $\sigma'_i(a|I) > \epsilon$, then there exists a $\delta > 0$ such that $q(z) > \delta$, thus ensuring Equation 6 is well-defined.

The algorithm works by sampling z using policy σ', storing $\pi^{\sigma'}(z)$. The single history is then traversed forward (to compute each player's probability of playing to reach each prefix of the history, $\pi^\sigma_i(h)$) and backward (to compute each player's probability of playing the remaining actions of the history, $\pi^\sigma_i(h, z)$). During the backward traversal, the sampled counterfactual regrets at each visited information set are computed (and added to the total regret).

$$\tilde{r}(I, a) = \begin{cases} w_I \cdot \left(1 - \sigma(a|z[I])\right) & \text{if } (z[I]a) \sqsubseteq z \\ -w_I \cdot \sigma(a|z[I]) & \text{otherwise} \end{cases} \text{, where } w_I = \frac{u_i(z) \pi^\sigma_{-i}(z) \pi^\sigma_i(z[I]a, z)}{\pi^{\sigma'}(z)} \quad (10)$$

One advantage of outcome-sampling MCCFR is that if our terminal history is sampled according to the opponent's policy, so $\sigma'_{-i} = \sigma_{-i}$, then the update no longer requires explicit knowledge of σ_{-i} as it cancels with the σ'_{-i}. So, w_I becomes $u_i(z)\pi^\sigma_i(z[I], z)/\pi^{\sigma'}_i(z)$. Therefore, we can use outcome-sampling MCCFR for online regret minimization. We would have to choose our own actions so that $\sigma'_i \approx \sigma^t_i$, but with some exploration to guarantee $q_j \geq \delta > 0$. By balancing the regret caused by exploration with the regret caused by a small δ (see Section 4 for how MCCFR's bound depends upon δ), we can bound the average overall regret as long as the number of playings T is known in advance. This effectively mimics the approach taking by Exp3 for regret minimization in normal-form games [9]. An alternative form for Equation 10 is recommended for implementation. This and other implementation details can be found in the appendix.

External-Sampling MCCFR. In *external-sampling MCCFR* we sample only the actions of the opponent and chance (those choices external to the player). We have a block $Q_\tau \in \mathcal{Q}$ for each pure strategy of the opponent and chance, *i.e.*,, for each deterministic mapping τ from $I \in \mathcal{I}_c \cup \mathcal{I}_{N\setminus\{i\}}$ to $A(I)$. The block probabilities are assigned based on the distributions f_c and σ_{-i}, so $q_\tau = \prod_{I\in\mathcal{I}_c} f_c(\tau(I)|I) \prod_{I\in\mathcal{I}_{N\setminus\{i\}}} \sigma_{-i}(\tau(I)|I)$. The block Q_τ then contains all terminal histories z consistent with τ, that is if ha is a prefix of z with $h \in I$ for some $I \in \mathcal{I}_{-i}$ then $\tau(I) = a$. In practice, we will not actually sample τ but rather sample the individual actions that make up τ only as needed. The key insight is that these block probabilities result in $q(z) = \pi^\sigma_{-i}(z)$. The algorithm iterates over $i \in N$ and for each doing a post-order depth-first traversal of the game tree, sampling actions at each history h where $P(h) \neq i$ (storing these choices so the same actions are sampled at all h in the same information set). Due to perfect recall it can never visit more than one history from the same information set during this traversal. For each such visited information set the sampled counterfactual regrets are computed (and added to the total regrets).

$$\tilde{r}(I,a) = (1 - \sigma(a|I)) \sum_{z\in Q\cap Z_I} u_i(z)\pi^\sigma_i(z[I]a, z) \tag{11}$$

Note that the summation can be easily computed during the traversal by always maintaining a weighted sum of the utilities of all terminal histories rooted at the current history.

4 Theoretical Analysis

We now present regret bounds for members of the MCCFR family, starting with an improved bound for vanilla CFR that depends more explicitly on the exact structure of the extensive game. Let $\vec{a}_i$ be a subsequence of a history such that it contains only player i's actions in that history, and let $\vec{A}_i$ be the set of all such player i action subsequences. Let $\mathcal{I}_i(\vec{a}_i)$ be the set of all information sets where player i's action sequence up to that information set is $\vec{a}_i$. Define the M-value for player i of the game to be $M_i = \sum_{\vec{a}_i\in\vec{A}_i} \sqrt{|\mathcal{I}_i(\vec{a})|}$. Note that $\sqrt{|\mathcal{I}_i|} \leq M_i \leq |\mathcal{I}_i|$ with both sides of this bound being realized by some game. We can strengthen vanilla CFR's regret bound using this constant, which also appears in the bounds for the MCCFR variants.

Theorem 3 *When using vanilla CFR for player i, $R_i^T \leq \Delta_{u,i} M_i \sqrt{|A_i|}/\sqrt{T}$.*

We now turn our attention to the MCCFR family of algorithms, for which we can provide probabilistic regret bounds. We begin with the most exciting result: showing that external-sampling requires only a constant factor more iterations than vanilla CFR (where the constant depends on the desired confidence in the bound).

Theorem 4 *For any $p \in (0,1]$, when using external-sampling MCCFR, with probability at least $1-p$, average overall regret is bounded by, $R_i^T \leq \left(1 + \frac{\sqrt{2}}{\sqrt{p}}\right) \Delta_{u,i} M_i \sqrt{|A_i|}/\sqrt{T}$.*

Although requiring the same order of iterations, note that external-sampling need only traverse a fraction of the tree on each iteration. For balanced games where players make roughly equal numbers of decisions, the iteration cost of external-sampling is $O(\sqrt{|H|})$, while vanilla CFR is $O(|H|)$, meaning external-sampling MCCFR requires asymptotically less time to compute an approximate equilibrium than vanilla CFR (and consequently chance-sampling CFR, which is identical to vanilla CFR in the absence of chance nodes).

Theorem 5 *For any $p \in (0,1]$, when using outcome-sampling MCCFR where $\forall z \in Z$ either $\pi^\sigma_{-i}(z) = 0$ or $q(z) \geq \delta > 0$ at every timestep, with probability $1-p$, average overall regret is bounded by $R_i^T \leq \left(1 + \frac{\sqrt{2}}{\sqrt{p}}\right)\left(\frac{1}{\delta}\right) \Delta_{u,i} M_i \sqrt{|A_i|}/\sqrt{T}$*

The proofs for the theorems in this section can be found in the appendix.

5 Experimental Results

We evaluate the performance of MCCFR compared to vanilla CFR on four different games. Goofspiel [10] is a bidding card game where players have a hand of cards numbered 1 to N, and take

Game	$\lvert H\rvert$ (10^6)	$\lvert\mathcal{I}\rvert$ (10^3)	l	M_1	M_2	t_{vc}	t_{os}	t_{es}
OCP	22.4	2	5	45	32	28s	46μs	99μs
Goof	98.3	3294	14	89884	89884	110s	150μs	150ms
LTTT	70.4	16039	18	1333630	1236660	38s	62μs	70ms
PAM	91.8	20	13	9541	2930	120s	85μs	28ms

Table 1: Game properties. The value of $|H|$ is in millions and $|\mathcal{I}|$ in thousands, and $l = \max_{h\in H}|h|$. t_{vc}, t_{os}, and t_{es} are the average wall-clock time per iteration[4] for vanilla CFR, outcome-sampling MCCFR, and external-sampling MCCFR.

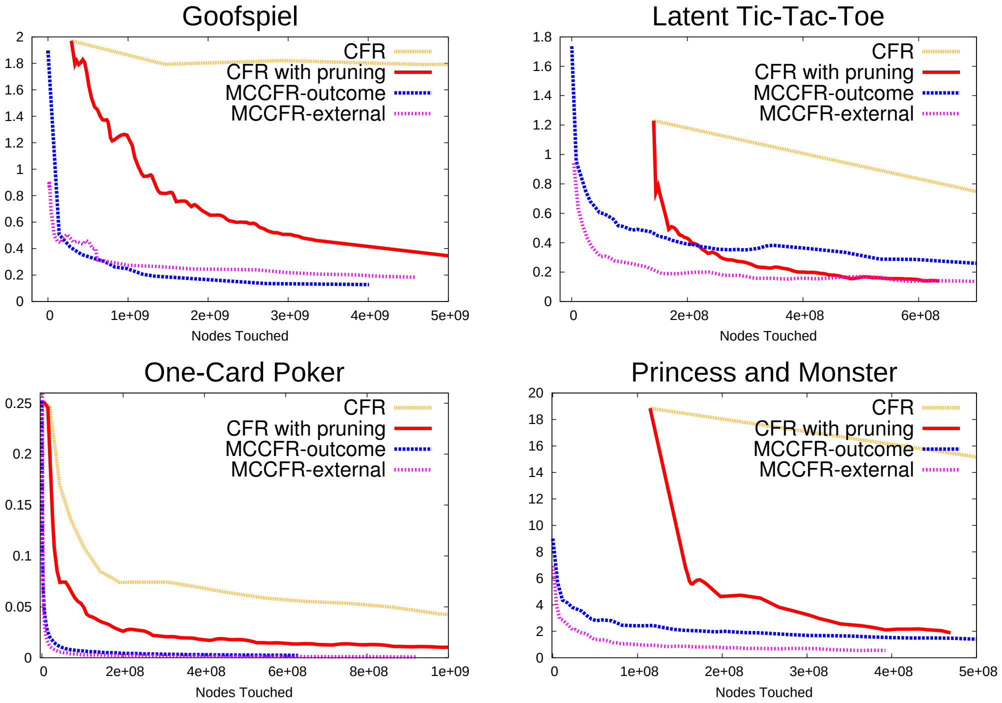

Figure 1: Convergence rates of Vanilla CFR, outcome-sampled MCCFR, and external-sampled MCCFR for various games. The y axis in each graph represents the exploitability of the strategies for the two players ϵ_σ (see Section 2.1).

turns secretly bidding on the top point-valued card in a point card stack using cards in their hands. Our version is less informational: players only find out the result of each bid and not which cards were used to bid, and the player with the highest total points wins. We use $N = 7$ in our experiments. One-Card Poker [11] is a generalization of Kuhn Poker [12], we use a deck of size 500. Princess and Monster [13, Research Problem 12.4.1] is a pursuit-evasion game on a graph, neither player ever knowing the location of the other. In our experiments we use random starting positions, a 4-connected 3 by 3 grid graph, and a horizon of 13 steps. The payoff to the evader is the number of steps uncaptured. Latent Tic-Tac-Toe is a twist on the classic game where moves are not disclosed until after the opponent's next move, and lost if invalid at the time they are revealed. While all of these games have imperfect information and roughly of similar size, they are a diverse set of games, varying both in the degree (the ratio of the number of information sets to the number of histories) and nature (whether due to chance or opponent actions) of imperfect information. The left columns of Table 1 show various constants, including the number of histories, information sets, game length, and M-values, for each of these domains.

We used outcome-sampling MCCFR, external-sampling MCCFR, and vanilla CFR to compute an approximate equilibrium in each of the four games. For outcome-sampling MCCFR we used an epsilon-greedy sampling profile σ'. At each information set, we sample an action uniformly ran-

[4] As measured on an 8-core Intel Xeon 2.5 GHz machine running Linux x86_64 kernel 2.6.27.

domly with probability ϵ and according to the player's current strategy σ^t. Through experimentation we found that $\epsilon = 0.6$ worked well across all games; this is interesting because the regret bound suggests δ should be as large as possible. This implies that putting some bias on the most likely outcome to occur is helpful. With vanilla CFR we used to an implementational trick called pruning to dramatically reduce the work done per iteration. When updating one player's regrets, if the other player has no probability of reaching the current history, the entire subtree at that history can be pruned for the current iteration, with no effect on the resulting computation. We also used vanilla CFR without pruning to see the effects of pruning in our domains.

Figure 1 shows the results of all four algorithms on all four domains, plotting approximation quality as a function of the number of nodes of the game tree the algorithm touched while computing. Nodes touched is an implementation-independent measure of computation; however, the results are nearly identical if total wall-clock time is used instead. Since the algorithms take radically different amounts of time per iteration, this comparison directly answers if the sampling variants' lower cost per iteration outweighs the required increase in the number of iterations. Furthermore, for any fixed game (and degree of confidence that the bound holds), the algorithms' average overall regret is falling at the same rate, $O(1/\sqrt{T})$, meaning that only their short-term rather than asymptotic performance will differ.

The graphs show that the MCCFR variants often dramatically outperform vanilla CFR. For example, in Goofspiel, both MCCFR variants require only a few million nodes to reach $\epsilon_\sigma < 0.5$ where CFR takes 2.5 billion nodes, three orders of magnitude more. In fact, external-sampling, which has the tightest theoretical computation-time bound, outperformed CFR and by considerable margins (excepting LTTT) in all of the games. Note that pruning is key to vanilla CFR being at all practical in these games. For example, in Latent Tic-Tac-Toe the first iteration of CFR touches 142 million nodes, but later iterations touch as few as 5 million nodes. This is because pruning is not possible in the first iteration. We believe this is due to dominated actions in the game. After one or two traversals, the players identify and eliminate dominated actions from their policies, allowing these subtrees to pruned. Finally, it is interesting to note that external-sampling was not uniformly the best choice, with outcome-sampling performing better in Goofspiel. With outcome-sampling performing worse than vanilla CFR in LTTT, this raises the question of what specific game properties might favor one algorithm over another and whether it might be possible to incorporate additional game specific constants into the bounds.

6 Conclusion

In this paper we defined a family of sample-based CFR algorithms for computing approximate equilibria in extensive games, subsuming all previous CFR variants. We also introduced two sampling schemes: outcome-sampling, which samples only a single history for each iteration, and external-sampling, which samples a deterministic strategy for the opponent and chance. In addition to presenting a tighter bound for vanilla CFR, we presented regret bounds for both sampling variants, which showed that external sampling with high probability gives an asymptotic computational time improvement over vanilla CFR. We then showed empirically in very different domains that the reduction in iteration time outweighs the increase in required iterations leading to faster convergence.

There are three interesting directions for future work. First, we would like to examine how the properties of the game effect the algorithms' convergence. Such an analysis could offer further algorithmic or theoretical improvements, as well as practical suggestions, such as how to choose a sampling policy in outcome-sampled MCCFR. Second, using outcome-sampled MCCFR as a general online regret minimizing technique in extensive games (when the opponents' strategy is not known or controlled) appears promising. It would be interesting to compare the approach, in terms of bounds, computation, and practical convergence, to Gordon's Lagrangian hedging [5]. Lastly, it seems like this work could be naturally extended to cases where we don't assume perfect recall. Imperfect recall could be used as a mechanism for abstraction over actions, where information sets are grouped by important partial sequences rather than their full sequences.

References

[1] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems 20 (NIPS)*, 2008.

[2] Andrew Gilpin, Samid Hoda, Javier Peña, and Tuomas Sandholm. Gradient-based algorithms for finding Nash equilibria in extensive form games. In *3rd International Workshop on Internet and Network Economics (WINE'07)*, 2007.

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

[4] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in game with incomplete information. Technical Report TR07-14, University of Alberta, 2007. `http://www.cs.ualberta.ca/research/techreports/2007/TR07-14.php`.

[5] Geoffrey J. Gordon. No-regret algorithms for online convex programs. In *In Neural Information Processing Systems 19*, 2007.

[6] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

[7] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, September 2000.

[8] D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

[9] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-arm bandit problem. In *36th Annual Symposium on Foundations of Computer Science*, pages 322–331, 1995.

[10] S. M. Ross. Goofspiel — the game of pure strategy. *Journal of Applied Probability*, 8(3):621–625, 1971.

[11] Geoffrey J. Gordon. No-regret algorithms for structured prediction problems. Technical Report CMU-CALD-05-112, Carnegie Mellon University, 2005.

[12] H. W. Kuhn. Simplified two-person poker. *Contributions to the Theory of Games*, 1:97–103, 1950.

[13] Rufus Isaacs. *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*. John Wiley & Sons, 1965.

A Appendix

This appendix first presents a detailed description of the MCCFR algorithm. We then give proofs to Theorems 3, 4, and 5. We begin with some preliminaries, then prove a general result about all members of the MCCFR family of algorithms (Theorem 7 in Section A.5). We then use that result to prove bounds for the MCCFR variants (Theorems 8 and 9 in Section A.6). We finally prove the tightened bound for vanilla CFR (Theorem 10 in Section A.7).

A.1 MCCFR Algorithm

The MCCFR algorithm is presented in detail in Algorithm 1.

Algorithm 1 Monte Carlo CFR with optimistic averaging

```
Require: a sampling scheme $\mathcal{S}$
Initialize information set markers: $\forall I, c_I \leftarrow 0$
Initialize regret tables: $\forall I, r_I[a] \leftarrow 0$.
Initialize cumulative strategy tables: $\forall I, s_I[a] \leftarrow 0$.
Initialize initial profile: $\sigma(I, a) \leftarrow 1/|A(I)|$
for $t = \{1, 2, 3, \cdots\}$ do
  for $i \in N$ do
    Sample a block of terminal histories $Q \in \mathcal{Q}$ using $\mathcal{S}$
    for each prefix history $z[I]$ of a terminal history $z \in Q$ with $P(z[I]) = i$ do
      for $a \in A(I)$ do
        Let $\tilde{r} = \tilde{r}(I, a)$, the sampled counterfactual regret
        $r_I[a] \leftarrow r_I[a] + \tilde{r}$
        $s_I[a] \leftarrow s_I[a] + (t - c_I)\pi_i^\sigma \sigma_i(I, a)$
      end for
      $c_I \leftarrow t$
      $\sigma_i \leftarrow$ RegretMatching($r_I$)
    end for
  end for
end for
```

In Algorithm 1, the average strategy is updated *optimistically* by weighting the update to the average strategy equally for every iteration not seen since the last time the information set was visited. Note: this can be corrected by maintaining weights at each parent information set which get updated whenever they are visited, and pushing the values of the weights down as needed (*lazy updating*). The average strategy can also be updated *stochastically* by weighting each update as the inverse of the probability of reaching the information set. The average strategy, $\bar{\sigma}$ is obtained by normalizing the values of the cumulative strategy tables s_I for each action at each information set I. Although *optimistic* averaging is not technically a correct average it performs well empirically.

We've discussed two novel sampling schemes in this work: *outcome-sampling* and *external sampling*.

A.1.1 Outcome Sampling

When using outcome-sampling, we can do the updates for each player simultaneously on a single pass over the one sampled terminal history. When $z[I]a$ is a prefix of z (action a was taken at I in our sampled history) then

$$\tilde{r}(I, a) = \tilde{v}_i(\sigma^t_{(I\to a)}, I) - \tilde{v}_i(\sigma^t, I) \tag{12}$$

$$= \frac{u_i(z)\pi^\sigma_{-i}(z[I])\pi^\sigma(z[I]a, z)}{\pi^{\sigma'}(z)} - \frac{u_i(z)\pi^\sigma_{-i}(z[I])\pi^\sigma(z[I], z)}{\pi^{\sigma'}(z)} \tag{13}$$

$$= \frac{u_i(z)\pi^\sigma_{-i}(z[I])}{\pi^{\sigma'}(z)}\left(\pi^\sigma(z[I]a, z) - \pi^\sigma(z[I], z)\right) \tag{14}$$

$$= W \cdot \left(\pi^\sigma(z[I]a, z) - \pi^\sigma(z[I], z)\right) \tag{15}$$

where

$$W = \frac{u_i(z)\pi^\sigma_{-i}(z[I])}{\pi^{\sigma'}(z)} \tag{16}$$

When $z[I]a$ is not a prefix of z, then $\tilde{v}_i(\sigma^t_{(I\to a)}, I) = 0$, so

$$\begin{aligned} \tilde{r}(I,a) &= 0 - \tilde{v}_i(\sigma^t, I) && (17)\\ &= -W \cdot \pi^\sigma(z[I], z) && (18) \end{aligned}$$

A.1.2 External Sampling

When using external sampling, we update for each player separately (one pass over the tree for each player). When updating I belonging to player i, note that $\pi^\sigma_{-i}(z[I], z) = \pi^\sigma_{-i}(z[I]a, z)$ since a is taken by i, not the opponent. Also note that $q(z) = \pi^\sigma_{-i}(z)$. We have the regret:

$$\begin{aligned} \tilde{r}(I,a) &= \sum_{z\in Q\cap Z_I} \left(\tilde{v}_i(\sigma^t_{(I\to a)}, I) - \tilde{v}_i(\sigma^t, I)\right) && (19)\\ &= \sum_{z\in Q\cap Z_I} \frac{u_i(z)\pi^\sigma_{-i}(z[I])}{q(z)} \left(\pi^\sigma(z[I]a,z) - \pi^\sigma(z[I],z)\right) && (20)\\ &= \sum_{z\in Q\cap Z_I} \frac{u_i(z)\pi^\sigma_{-i}(z[I])\pi^\sigma_{-i}(z[I]a,z)}{q(z)} \left(\frac{\pi^\sigma(z[I]a,z) - \pi^\sigma(z[I],z)}{\pi^\sigma_{-i}(z[I]a,z)}\right) && (21)\\ &= \sum_{z\in Q\cap Z_I} \frac{u_i(z)\pi^\sigma_{-i}(z)}{q(z)} \left(\pi^\sigma_i(z[I]a,z) - \pi^\sigma_i(z[I],z)\right) && (22)\\ &= \sum_{z\in Q\cap Z_I} u_i(z)\pi^\sigma_i(z[I]a,z)\left(1-\sigma(a|I)\right) && (23)\\ &= \left(1-\sigma(a|I)\right) \sum_{z\in Q\cap Z_I} u_i(z)\pi^\sigma_i(z[I]a,z) && (24)\\ &= \frac{\sigma(a|I)}{\sigma(a|I)}\left(1-\sigma(a|I)\right) \sum_{z\in Q\cap Z_I} u_i(z)\pi^\sigma_i(z[I]a,z) && (25)\\ &= \left(\frac{1}{\sigma(a|I)} - 1\right) \sum_{z\in Q\cap Z_I} u_i(z)\pi^\sigma_i(z[I],z) && (26)\\ &= \frac{1}{\sigma(a|I)} \sum_{z\in Q\cap Z_I} u_i(z)\pi^\sigma_i(z[I],z) - \sum_{z\in Q\cap Z_I} u_i(z)\pi^\sigma_i(z[I],z) && (27) \end{aligned}$$

The sum is the expected utility to player i from $z[I]$, assuming the opponent plays with the deterministic mapping τ that was sampled from their mixed strategy. Here, the left-side term represents the expected utility if player i chooses a at $z[I]$ and then the players continue with their strategies afterwards and the right-side term represents the expected utility if player i plays according to σ at $z[I]$. In practice the left-side term is computed by a tree traversal for each action taken from $z[I]$ and then the right-side sum is computed as a weighted sum of these resulting expected utilities.

A.2 Preliminaries

There are several basic properties of random variables and real numbers that are necessary to prove the main results.

Lemma 2 *For any random variable X:*

$$\Pr[|X| \geq k\sqrt{\mathbf{E}[X^2]}] \leq \frac{1}{k^2}. \tag{28}$$

Proof: Markov's Inequality states, if Y is always non-negative:

$$\Pr[Y \geq j\mathbf{E}[Y]] \leq \frac{1}{j}. \tag{29}$$

By setting $Y = X^2$:

$$\Pr[X^2 \geq j\mathbf{E}[X^2]] \leq \frac{1}{j} \tag{30}$$

$$\Pr[|X| \geq \sqrt{j\mathbf{E}[X^2]}] \leq \frac{1}{j}. \tag{31}$$

Replacing $k = \sqrt{j}$:

$$\Pr[|X| \geq k\sqrt{\mathbf{E}[X^2]}] \leq \frac{1}{j^2}. \tag{32}$$

■

Lemma 3 *If $a_1 \ldots, a_n$ are non-negative real numbers in the interval $[0, 1]$ where $\sum_{i=1}^{n} a_i = S$, then $\sum_{i=1}^{n}(a_i)^2 \leq S$.*

Proof: Assume without loss of generality that $n \geq \lceil S \rceil$..

Suppose that there are two elements a_i,a_j, where $a_i < 1$ and $a_j < 1$. If $a_i + a_j \leq 1$, then:

$$(a_i)^2 + (a_j)^2 \leq (a_i)^2 + 2a_ia_j + (a_j)^2 \tag{33}$$

$$\leq (a_i + a_j)^2. \tag{34}$$

Thus, it is better to have $(a_i + a_j, 0)$. If $a_i + a_j > 1$, then define $A = a_i + a_j$, and define $f(x) = (A - x)^2 + x^2$. Setting the derivative to zero:

$$0 = f'(x) \tag{35}$$

$$f'(x) = -2(A - x) + 2x \tag{36}$$

$$2A = 4x \tag{37}$$

$$\frac{A}{2} = x \tag{38}$$

Upon further observation, $f''(x) = 4$, implying that $\frac{A}{2}$ is a minimal point. Therefore, since the critical points of $f(x)$ are $\frac{A}{2}$ and the limits of the feasible region, namely $A - 1$, and 1, then the limits of the feasible region must be the maximal points.

Therefore, for any two a_i and a_j, either:

1. One or the other is zero, or:

2. One is equal to the other.

Therefore, there can be no more than one i such that $a_i \in (0, 1)$, all others must be equal to zero or one. Define $i^* = \lfloor S \rfloor$. Without loss of generality, assume for all $i \in \{1 \ldots i^*\}$, $a_i = 1$, $a_{i^*+1} = S - \lfloor S \rfloor$, and for all $i \in \{i^* + 2 \ldots n\}$, $a_i = 0$. The result follows directly. ■

Lemma 4 *If $a_1 \ldots, a_n$ are non-negative real numbers where $\sum_{i=1}^{n} a_i = S$, then $\sum_{i=1}^{n} \sqrt{a_i} \leq \sqrt{Sn}$.*

Proof: We prove this by induction on n. If $n = 1$, then the result is trivial. Otherwise, define $x = \sum_{i=1}^{n} a_i$, so that $a_n + x = S$, and therefore by induction $\sum_{i=1}^{n} \sqrt{a_i} \leq \sqrt{x(n-1)} + \sqrt{S - x}$. Define $f(x) = \sqrt{x(n-1)} + \sqrt{S - x}$. To maximize $f(x)$, we observe that 0 and S are critical points, and we take the derivative and set it to zero:

$$f'(x) = 0 \tag{39}$$

$$f'(x) = \frac{0.5(n-1)}{\sqrt{x(n-1)}} - \frac{0.5}{\sqrt{S-x}} \tag{40}$$

$$\frac{0.5\sqrt{n-1}}{\sqrt{x}} = \frac{0.5}{\sqrt{S-x}} \tag{41}$$

$$\frac{x}{n-1} = S - x \tag{42}$$

$$x\left(1 + \frac{1}{n-1}\right) = S \tag{43}$$

$$x\left(\frac{n-1+1}{n-1}\right) = S \tag{44}$$

$$x = \frac{S(n-1)}{n} \tag{45}$$

Therefore, substituting the three critical points yields:

$$f(0) = \sqrt{S} \tag{46}$$

$$f(S) = \sqrt{S(n-1)} \tag{47}$$

$$f\left(\frac{S(n-1)}{n}\right) = \sqrt{\frac{S(n-1)(n-1)}{n}} + \sqrt{S - \frac{S(n-1)}{n}} \tag{48}$$

$$= (n-1)\sqrt{\frac{S}{n}} + \sqrt{\frac{S}{n}} \tag{49}$$

$$= \sqrt{Sn} \tag{50}$$

The maximum of these is $\sqrt{Sn}$, establishing the inductive step. ■

Lemma 5 *If $b_1 \ldots, b_n$ are non-negative real numbers where $\sum_{i=1}^n b_i^2 = S$, then $\sum_{i=1}^n b_i \leq \sqrt{Sn}$.*

Proof: Let $a_i = b_i^2$ and apply Lemma 4. ■

Lemma 6 *Given nonnegative reals $a_{i,j}$ in $[0,1]$, where $\sum_{i=1}^m \sum_{j=1}^n a_{m,n} = S$, then:*

$$\sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} (a_{m,n})^2} \leq \sqrt{mS}. \tag{51}$$

A.3 Blackwell's Approachability Theorem

Consider the following more sophisticated bound for the regret matching procedure using Blackwell's approachability.

Lemma 7 *For all real a, define $a^+ = \max(a, 0)$. For all a, b, it is the case that*

$$\left((a+b)^+\right)^2 \leq (a^+)^2 + 2(a^+)b + b^2 \tag{52}$$

Proof: We prove this by enumerating the possibilities:

1. $a \leq 0$. Then $a^+ = 0$, so we have:

$$\left((a+b)^+\right)^2 \leq (b^+)^2 \tag{53}$$

$$\leq b^2, \tag{54}$$

and:

$$(a^+)^2 + 2(a^+)b + b^2 = b^2. \tag{55}$$

2. $a \geq 0$,$b \geq -a$. Then $a = a^+$ and $(a+b)^+ = (a+b)$. So:

$$\left((a+b)^+\right)^2 = (a+b)^2. \tag{56}$$

Also:

$$\begin{aligned}(a^+)^2 + 2(a^+)b + b^2 &= a^2 + 2ab + b^2 && (57)\\ &= (a+b)^2 && (58)\end{aligned}$$

3. $a \geq 0$, $b \leq -a$. Then $a = a^+$, and $(a+b)^+ = 0$. So:

$$\left((a+b)^+\right)^2 = 0. \tag{59}$$

Also:

$$\begin{aligned}(a^+)^2 + 2(a^+)b + b^2 &= a^2 + 2ab + b^2 && (60)\\ &= (a+b)^2 && (61)\\ &\geq 0 && (62)\end{aligned}$$

■

Define $R^+_{\sum,T} = \sum_{a \in A} R^+_T(a)$. Regret matching is a strategy σ_{T+1} where:

$$\sigma_{T+1}(a) = \begin{cases} \frac{R^+_T(a)}{R^+_{\sum,T}} & \text{if } R^+_{\sum,T} > 0 \\ \frac{1}{|A|} & \text{otherwise} \end{cases} \tag{63}$$

Lemma 8 *If regret matching is used, then:*

$$\sum_{a \in A} R^+_T(a) r_{T+1}(a) \leq 0 \tag{64}$$

Proof: If $R^+_{\sum,T} \leq 0$, then for all $a \in A$, $R^+_T(a) = 0$, and the result is trivial. Otherwise:

$$\begin{aligned}\sum_{a \in A} R^+_T(a) r_{T+1}(a) &= \sum_{a \in A} R^+_T(a)(u_{T+1}(a) - u_{T+1}(\sigma_t)) && (65)\\ &= \left(\sum_{a \in A} R^+_T(a) u_{T+1}(a)\right) - \left(u_{T+1}(\sigma_t) \sum_{a \in A} R^+_T(a)\right) && (66)\\ &= \left(\sum_{a \in A} R^+_T(a) u_{T+1}(a)\right) - \left(\sum_{a' \in A} \sigma_{T+1}(a') u_{T+1}(a')\right) R^+_{\sum,T} && (67)\\ &= \left(\sum_{a \in A} R^+_T(a) u_{T+1}(a)\right) - \left(\sum_{a' \in A} \frac{R^+_T(a')}{R^+_{\sum,T}} u_{T+1}(a')\right) R^+_{\sum,T} && (68)\\ &= \left(\sum_{a \in A} R^+_T(a) u_{T+1}(a)\right) - \left(\sum_{a' \in A} R^+_T(a') u_{T+1}(a')\right) && (69)\\ &= 0 && (70)\end{aligned}$$

■

Theorem 6 *Define* Δ_t *to be* $\max_{a,a' \in A}(u_t(a) - u_t(a'))$. *Then regret matching yields:*

$$\sum_{a \in A} (R^+_T(a))^2 \leq \frac{1}{T^2} \sum_{t=1}^{T} |A| (\Delta_t)^2. \tag{71}$$

Proof: We prove this by recursion on T. The base case (for $T = 1$) is obvious. Assuming this holds for $T-1$, we prove it holds for T. Since $R_T(a) = \frac{(T-1)}{T} R_{T-1}(a) + \frac{1}{T} r_T(a)$, by Lemma 7:

$$(R_T^+(a))^2 \leq (\frac{(T-1)R_{T-1}^+(a)}{T})^2 + 2\frac{T-1}{T^2} R_{T-1}^+(a) r_T(a) + (\frac{r_T(a)}{T})^2 \tag{72}$$

Summing yields:

$$\sum_{a\in A}(R_T^+(a))^2 \leq \sum_{a\in A}\left(\left(\frac{T-1}{T}\right)^2 (R_{T-1}^+(a))^2 + 2\frac{T-1}{T^2} R_{T-1}^+(a) r_T(a) + \frac{1}{T^2}(r_T(a))^2\right) \tag{73}$$

By Lemma 8, $\sum_{a\in A} R_{T-1}^+(a) r_T(a) = 0$, so:

$$\sum_{a\in A}(R_T^+(a))^2 \leq \left(\left(\frac{T-1}{T}\right)^2 \sum_{a\in A}(R_{T-1}^+(a))^2\right) + \left(\frac{1}{T^2}\sum_{a\in A}(r_T(a))^2\right) \tag{74}$$

By induction:

$$\sum_{a\in A}(R_{T-1}^+(a))^2 \leq \frac{1}{(T-1)^2}\sum_{t=1}^{T-1} |A|(\Delta_t)^2. \tag{75}$$

Note that $|r_T(a)| \leq \Delta_T$. So:

$$\sum_{a\in A}(R_T^+(a))^2 \leq \frac{1}{T^2}\left(\sum_{t=1}^{T-1} |A|(\Delta_t)^2\right) + |A|(\Delta_T)^2. \tag{76}$$

■

A.4 Deterministic Strategies

Before delving into the general proof, we need a few gory details involving deterministic strategies.

A **deterministic strategy** $\sigma_i : \mathcal{I}_i \to A(i)$ maps each information set $I_i \in \mathcal{I}_i$ to an action $a \in A(I_i)$. Define $\widehat{\Sigma}_i$ to be the set of deterministic strategies for i, and $\widehat{\Sigma} = \prod_{i\in N'} \widehat{\Sigma}_i$, and $\widehat{\Sigma}_{-i} = \prod_{j\in N'\setminus i} \widehat{\Sigma}_j$.

Define $I(h)$ to be the information set $I_i \in \mathcal{I}_{P(h)}$ containing h. Given a deterministic strategy profile σ, we can make it into a function from a history to the next action, defined as $\sigma(h) = \sigma_{P(h)}(I(h))$. The terminal history $h(\sigma)$ is the unique $h \in Z$ such that, for all $t \in \{0 \ldots |h|-1\}$, $\sigma(h(t)) = h_{t+1}$. An information set I is **reached with $\boldsymbol{\sigma}$** if for some $h' \sqsubseteq h(\sigma)$, $h' \in I$. In a game with perfect recall, define $h(\sigma, I)$ to be the unique $h' \in I$ where $h' \sqsubseteq h(\sigma)$.

If no deterministic strategy σ_i of i allows I to be reached with (σ_{-i}, σ_i), then I is **unreachable with $\boldsymbol{\sigma_{-i}}$**.

In a game with perfect recall, given σ_{-i}, each information set $I_i \in \mathcal{I}_i$, given two deterministic strategies σ_i and σ_i', if σ_i and σ_i' both reach I, then $h((\sigma_{-i}, \sigma_i), I) = h((\sigma_{-i}, \sigma_i'), I)$. Therefore, if I is reachable with σ_{-i} we define $h(\sigma_{-i}, I) = h((\sigma_{-i}, \sigma_i), I)$ for some σ_i such that I is reached with (σ_{-i}, σ_i). In general, for any set $S \subseteq N'$, I is reachable with $\sigma_S = \{\hat{\sigma}_i\}_{i\in S}$ if there exists a set $\sigma_{N'\setminus S} = \{\hat{\sigma}_i\}_{i\in N'\setminus S}$ such that I is reachable with $(\sigma_S, \sigma_{N'\setminus S})$.

Given a history $h' \in H$, one can consider what would happen if σ was used to play h' to termination. In particular, define $h(\sigma, h') \in Z$ to be the unique history $h \in Z$ such that $h' \sqsubseteq h$ and for all $t \in \{|h'| \ldots |h|-1\}$, $\sigma(h(t)) = h_{t+1}$. Thus, for all $h \in H$, we can define $u_i(h', \sigma) = u_i(h(\sigma, h'))$.

Given $\vec{a}$, σ_i **obliviously plays $\vec{\boldsymbol{a}}$** if the strategy that plays the actions in $\vec{a}$ deterministically in sequence. In particular, for any information set $I_i \in \mathcal{I}_i$, define $c(I_i) = |X_i(h)|$, the length of the sequence of information sets and actions reached by this player before this information set, for any $h \in I_i$ (in a game with perfect recall, this is well-defined). Therefore, $\sigma_i(I_i) = \vec{a}_{c(I_i)+1}$, or is arbitrary if $c(I_i)+1$ is greater than the number of elements of $\vec{a}$.

Lemma 9 *For any deterministic profile σ_{-i}, for any $\vec{a}$, if $I_i \in \mathcal{I}_i(\vec{a})$ is reachable with σ_{-i}, then it is reachable with (σ_{-i}, σ_i), where σ_i obliviously plays $\vec{a}$.*

Proof: Since I_i is reachable with σ_{-i}, then there exists some σ_i such that I_i is reachable with (σ_{-i}, σ_i). By definition, the history $h(\sigma_{-i}, \sigma_i)$ has a prefix $h' \in I_i$. Define $\vec{a}(t)$ to be the first t elements of $\vec{a}$, and define σ_i^t to be the strategy that obliviously plays $\vec{a}(t)$, and arbitrary decisions are equal to σ_i. We will prove by recursion on t that for all $t \leq \vec{a}$, $h(\sigma_{-i}, \sigma_i^t) = h(\sigma_{-i}, \sigma_i)$. First of all observe that $\sigma_i^0 = \sigma_i$, so the basis of the recursion holds. For the inductive step, we assume that $h(\sigma_{-i}, \sigma_i^{t-1}) = h(\sigma_{-i}, \sigma_i)$, and try to prove that $h(\sigma_{-i}, \sigma_i^t) = h(\sigma_{-i}, \sigma_i)$. Since $h' \in I_i$, then $X(h') = ((I_1, a_1) \dots (I_k, a_k))$, and since $I_i \in \mathcal{I}_i(\vec{a})$, then $\vec{a} = (a_1, \dots, a_k)$. Therefore, define h'' to be the prefix of h' in I_t. Note that σ_i^{t-1} is in control for all $I_1 \dots I_{t-1}$, and then σ_i selects a_t in information set I_t. However, since σ_t would have also selected a_t by definition, then $h(\sigma_{-i}, \sigma_i^t) = h(\sigma_{-i}, \sigma_i^{t-1}) = h(\sigma_{-i}, \sigma_i)$. Note that changing later actions of σ_i does not affect whether or not h' is played, so that any arbitrary deterministic strategy which is $\vec{a}$ oblivious will work. ■

Lemma 10 *For any deterministic profile σ_{-i}, for any $\vec{a}$, there is no more than one reachable $I_i \in \mathcal{I}_i(\vec{a})$.*

Proof: Consider two information sets $I_i', I_i'' \in \mathcal{I}_i(\vec{a})$ where $I_i' \neq I_i''$, and for the sake of contradiction, assume both are reachable with σ_{-i}. Given a σ_i which is $\vec{a}$-oblivious, then I_i' and I_i'' are both reachable with (σ_{-i}, σ_i). But there is only one history generated, $h(\sigma_{-i}, \sigma_i)$, and therefore there must exist $h' \in I_i'$ and $h'' \in I_i''$, both prefixes of $h(\sigma_{-i}, \sigma_i)$. But that implies that $h' \sqsubseteq h''$ or vice-versa, meaning that in a perfect recall game, the sequence of prior information sets and actions of either I_i' or I_i'' must include the other, an obvious contradiction to them both having action sequences of equal size. ■

Lemma 11 *For any strategy $\sigma_j \in \Sigma_j$, there exists a distribution $\rho \in \Delta(\hat{\Sigma}_j)$ such that for any $h \in H$,*

$$\Pr_{\hat{\sigma}_j \in \rho_j}[\forall (I, a) \in X_j(h), \hat{\sigma}_j(I) = a] = \pi^{\sigma_j}(h). \tag{77}$$

Proof: First, we define $\rho(\hat{\sigma}_j)$ to be:

$$\rho(\hat{\sigma}_j) = \prod_{I \in \mathcal{I}_j} \sigma_j(I)(\hat{\sigma}_j(I)). \tag{78}$$

In other words, the probability of playing $\hat{\sigma}_j$ is the probability of playing like $\hat{\sigma}_j$ everywhere. Summing over all actions outside of $X_j(h)$ gives the lemma. ■

Lemma 12 *Given a single player strategy $\sigma_i \in \Sigma_i$, and ρ generated as in Lemma 11, then:*

$$\Pr_{\hat{\sigma}_i \in \rho}[\mathrm{Reach}_i^{\hat{\sigma}_i}(h)] = \pi_i^{\sigma_i}(h). \tag{79}$$

Lemma 13 *Given a history $h \in H$, $\mathrm{Reach}_i^{\hat{\sigma}_i}(h)$ if and only if for all $(I, a) \in X_i(h)$, $\hat{\sigma}_i(I) = a$.*

Proof: First, if for all $(I, a) \in X_i(h)$, $\hat{\sigma}_i(I) = a$, then we can define $\hat{\sigma}_j$ such that for all $(I, a) \in X_j(h)$, $\hat{\sigma}_j(I) = a$, and for all $I \notin X_j(h)$, set $\hat{\sigma}_j(I)$ to be arbitrary. Note that for all $h' \sqsubseteq h$, there exists an $(I, a) \in X_{P(h')}(h)$ where $h' \in I$ and $h_{|h'|+1} = a$. Therefore, for all $h' \sqsubseteq h$, $\hat{\sigma}_{P(h')}(I) = h_{|h'|+1}$, implying that $\hat{\sigma}$ reaches h.

If $\mathrm{Reach}_i^{\hat{\sigma}_i}(h)$, then there exists a $\hat{\sigma}_{-i}$ such that $h \sqsubseteq h(\hat{\sigma}_i, \hat{\sigma}_{-i})$. Therefore, for all $h' \sqsubseteq h$, $\hat{sigma}(h') = h_{|h'|+1}$, and $\hat{\sigma}_{P(h')}(I(h')) = h_{|h'|+1}$. For all $(I, a) \in X_i(h)$, $I = I(h'')$ and $a = h_{|h''|+1}$ for some $h'' \sqsubseteq h$. Moreover, $P(h'') = i$, implying that $\hat{\sigma}_i(I(h'')) = h_{|h''|+1}$. ■

Corollary 1 *Given any $h \in H$, given $i \in N'$, there exists a $\hat{\sigma}_i \in \widehat{\Sigma}_i$ that reaches h.*

Proof: For any history h, it is easy to construct a strategy which satisfies Lemma 13. ■

Lemma 14 *Given a set of strategies $\hat{\sigma}_S = \{\hat{\sigma}_i\}_{i \in S}$, if for all $i \in S$, $\mathrm{Reach}_i^{\hat{\sigma}_i}(h)$, then $\mathrm{Reach}_S^{\hat{\sigma}_S}(h)$.*

Proof: By Corollary 1, for every $i \in N' \backslash S$, there exists a strategy $\hat{\sigma}_i$ that reaches h. By Lemma 13, for all $i \in N'$, for all $(I, a) \in X_i(h)$, $\hat{\sigma}_i(I) = a$. Moreover, this implies that these strategies reconstruct h. ■

Lemma 15 *For any strategy profile $\sigma_{-i} \in \Sigma_{-i}$, there exists a distribution $\rho \in \Delta(\widehat{\Sigma}_{-i})$ such that for all $I \in \mathcal{I}_i$, $\pi_{-i}^{\sigma_{-i}}(I) = \sum_{\hat{\sigma}_{-i} \in \widehat{\Sigma}_{-i}:\text{Reach}(\hat{\sigma}_{-i}, I)} \rho(\hat{\sigma}_{-i})$.*

Proof: For all $j \in N' \backslash i$, using Lemma 11, we generate a strategy $\rho_j \in \Delta(\widehat{\Sigma}_j)$. Define ρ to be the distribution over $\Delta(\widehat{\Sigma}_{-i})$ obtained by independently sampling each $\hat{\sigma}_j$ by ρ_j; formally,

$$\rho(\hat{\sigma}_{-i}) = \prod_{j \in N' \backslash i} \rho_j(\hat{\sigma}_j). \tag{80}$$

Consider a history $h \in H$. By Lemma 12, $\pi_j^{\sigma_j}(h) = \Pr_{\hat{\sigma}_j \in \rho_j}[\text{Reach}_j^{\hat{\sigma}_j}(h)]$. Since the strategies are selected independently:

$$\Pr_{\hat{\sigma}_{-i} \in \rho}[\forall j \in N' \backslash i, \text{Reach}_j^{\hat{\sigma}_j}(h)] = \prod_{j \in N' \backslash i} \Pr_{\hat{\sigma}_j \in \rho_j}[\text{Reach}_j^{\hat{\sigma}_j}(h)] \tag{81}$$

$$= \prod_{j \in N' \backslash i} \pi^{\hat{\sigma}_j}(h) \tag{82}$$

$$= \pi^{\hat{\sigma}_{-i}}(h) \tag{83}$$

If we sum over all $h \in I$, we get the result. ■

Lemma 16 *For any strategy profile σ_{-i}, for any $\vec{a}$:*

$$\sum_{I \in \mathcal{I}_i(\vec{a})} \pi_{-i}^{\sigma_{-i}}(I) \leq 1 \tag{84}$$

Proof: This follows directly from Lemma 15 and Lemma 10. ■

A.5 General MCCFR Bound

We begin by proving a very general bound applicable to all algorithms in the MCCFR family. First, define $\mathcal{B}_i = \{\mathcal{I}_i(\vec{a}) : \vec{a} \in \vec{A}_i\}$, so $M = \sum_{B \in \mathcal{B}_i} \sqrt{|B|}$.

Theorem 7 *For any $p \in (0, 1]$, when using any algorithm in the MCCFR family such that for all $Q \in \mathcal{Q}$ and $B \in \mathcal{B}$,*

$$\sum_{I \in B} \left(\sum_{z \in Q \cap Z_I} \frac{\pi^\sigma(z[I], z)\pi_{-i}^\sigma(z[I])}{q(z)} \right)^2 \leq \frac{1}{\delta^2} \tag{85}$$

where $\delta \leq 1$, then with with probability at least $1 - p$, average overall regret is bounded by,

$$R_i^T \leq \left(1 + \frac{2}{\sqrt{p}}\right)\left(\frac{1}{\delta}\right) \frac{\Delta_{u,i} M_i \sqrt{|A_i|}}{\sqrt{T}}. \tag{86}$$

Proof: Define $r_i^t(I, a)$ to be the unsampled immediate counterfactual regret and $\tilde{r}_i^t(I, a)$ to be the sampled immediate counterfactual regret. Formally,

$$r_i^t(I, a) = \left(v_i(\sigma_{(I \to a)}^t, I) - v_i(\sigma^t, I)\right) \tag{87}$$

$$\tilde{r}_i^t(I, a) = \left(\tilde{v}_i(\sigma_{(I \to a)}^t, I) - \tilde{v}_i(\sigma^t, I)\right) \tag{88}$$

$$R_i^T(I) = \frac{1}{T} \max_{a \in A(I)} \sum_{t=1}^T r_i^t(I, a) \tag{89}$$

$$\tilde{R}_i^T(I) = \frac{1}{T} \max_{a \in A(I)} \sum_{t=1}^T \tilde{r}_i^t(I, a) \tag{90}$$

Let $Q_t \in \mathcal{Q}$ be the block sampled at time t. Note that we can bound the difference between two sampled counterfactual values for information set I at time t by,

$$\left(\tilde{v}_i(\sigma^t_{(I\to a)}, I) - \tilde{v}_i(\sigma^t, I)\right) \le \Delta^t_{u,i}(I) \equiv \Delta_{u,i} \sum_{z \in Q_t \cap Z_I} \frac{\pi^\sigma(z[I], z)\pi^\sigma_{-i}(z[I])}{q(z)} \tag{91}$$

so by our assumption,

$$\sum_{I \in B} \Delta^t_{u,i}(I)^2 \le \frac{\Delta^2_{u,i}}{\delta^2} \tag{92}$$

So we can apply Theorem 6, to get,

$$\tilde{R}^T_i(I) \le \frac{\sqrt{|A(I)| \sum_{t=1}^T (\Delta^t_{u,i}(I))^2}}{T} \tag{93}$$

Using Lemma 5,

$$\begin{aligned}
\sum_{I \in B} \tilde{R}^T_i(I) &\le \frac{\sqrt{|B||A(B)| \sum_{I \in B} \sum_{t=1}^T (\Delta^t_{u,i}(I))^2}}{T} && (94)\\
&\le \frac{\sqrt{|B||A(B)| \sum_{t=1}^T \sum_{I \in B} (\Delta^t_{u,i}(I))^2}}{T} && (95)\\
&\le \frac{\sqrt{|B||A(B)| \sum_{t=1}^T \Delta^2_{u,i}/\delta^2}}{T} && (96)\\
&\le \frac{\Delta_{u,i}\sqrt{|B||A(B)|}}{\delta\sqrt{T}} && (97)
\end{aligned}$$

The average overall sampled regret then can be bounded by,

$$\begin{aligned}
\tilde{R}^T_i &\le \sum_{B \in \mathcal{B}_i} \sum_{I \in B} \tilde{R}^T_i(I) && (98)\\
&\le \sum_{B \in \mathcal{B}_i} \frac{\Delta_{u,i}\sqrt{|B||A(B)|}}{\delta\sqrt{T}} && (99)\\
&\le \frac{\Delta_{u,i}\sqrt{|A_i|}\sum_{B \in \mathcal{B}_i}\sqrt{|B|}}{\delta\sqrt{T}} && (100)\\
&\le \frac{\Delta_{u,i} M_i \sqrt{|A_i|}}{\delta\sqrt{T}} && (101)
\end{aligned}$$

We now need to prove that R and $\tilde{R}$ are similar. This last portion is tricky. Since the algorithm is randomized, we cannot guarantee that every information set is reached, let alone that it has converged. Therefore, instead of proving a bound on the absolute difference of R and $\tilde{R}$, we focus on proving a probabilistic connection.

In particular, we will bound the expected squared difference between $\sum_{I \in \mathcal{I}_i} R^T_i(I)$ and $\sum_{I \in \mathcal{I}_i} \tilde{R}^T_i(I)$ in order to prove that they are close, and then use Lemma 2 to bound the absolute value. We begin by focusing on the similarity of the counterfactual regret ($R^T_i(I)$ and $\tilde{R}^T_i(I)$) in every node, by focusing on the similarity of the counterfactual regret of a particular action at a particular time ($r^t_i(I, a)$ and $\tilde{r}^t_i(I, a)$). By the Lemma from the main paper, we know that $\mathbf{E}[r^t_i(I, a) - \tilde{r}^t_i(I, a)] = 0$.

From Lemma 5 we have,

$$\mathbf{E}\left[\left(\sum_{I \in \mathcal{I}_i} (R^T_i(I) - \tilde{R}^T_i(I))\right)^2\right] \le |\mathcal{I}_i| \sum_{I \in \mathcal{I}_i} \mathbf{E}\left[(R^T_i(I) - \tilde{R}^T_i(I))^2\right] \tag{102}$$

So,

$$(R_i^T(I) - \tilde{R}_i^T(I))^2 = \left(\frac{1}{T}\max_{a\in A(I)}\sum_{t=1}^T r_i^t(I,a) - \frac{1}{T}\max_{a\in A(I)}\sum_{t=1}^T \tilde{r}_i^t(I,a)\right)^2 \tag{103}$$

$$(R_i^T(I) - \tilde{R}_i^T(I))^2 \leq \frac{1}{T^2}\left(\max_{a\in A(I)}\left(\sum_{t=1}^T r_i^t(I,a) - \sum_{t=1}^T \tilde{r}_i^t(I,a)\right)\right)^2 \tag{104}$$

$$(R_i^T(I) - \tilde{R}_i^T(I))^2 \leq \frac{1}{T^2}\left(\max_{a\in A(I)}\left(\sum_{t=1}^T \left|r_i^t(I,a) - \tilde{r}_i^t(I,a)\right|\right)\right)^2 \tag{105}$$

Note that if $f(x)$ is monotonically increasing on the non-negative numbers, then $f(\max_a |x_a|) = \max_a f(|x_a|)$.

$$(R_i^T(I) - \tilde{R}_i^T(I))^2 \leq \frac{1}{T^2}\max_{a\in A(I)}\left(\sum_{t=1}^T r_i^t(I,a) - \sum_{t=1}^T \tilde{r}_i^t(I,a)\right)^2 \tag{106}$$

$$(R_i^T(I) - \tilde{R}_i^T(I))^2 \leq \frac{1}{T^2}\sum_{a\in A(I)}\left(\sum_{t=1}^T r_i^t(I,a) - \sum_{t=1}^T \tilde{r}_i^t(I,a)\right)^2 \tag{107}$$

$$\mathbf{E}[(R_i^T(I) - \tilde{R}_i^T(I))^2] \leq \frac{1}{T^2}\sum_{a\in A(I)}\sum_{t=1}^T \mathbf{E}[\left(r_i^t(I,a) - \tilde{r}_i^t(I,a)\right)^2] \tag{108}$$

The final step is because if $t \neq t'$, then $\mathbf{E}[(r_i^t(I,a) - \tilde{r}_i^t(I,a))(r_i^{t'}(I,a) - \tilde{r}_i^{t'}(I,a))] = 0$, because if $t > t'$, then after time t', $\tilde{r}_t(I,a)$ is an unbiased estimator of $r_i^t(I,a)$ (and vice-versa). Substituting back into Equation 102:

$$\mathbf{E}\left[(\sum_{I\in\mathcal{I}_i}(R_i^T(I) - \tilde{R}_i^T(I)))^2\right] \leq \frac{|\mathcal{I}_i|}{T^2}\sum_{I\in\mathcal{I}_i}\sum_{a\in A(I)}\sum_{t=1}^T \mathbf{E}\left[\left(r_i^t(I,a) - \tilde{r}_i^t(I,a)\right)^2\right] \tag{109}$$

$$\leq \frac{|\mathcal{I}_i|}{T^2}\sum_{t=1}^T\sum_{B\in\mathcal{B}_i}\sum_{a\in A(B)}\sum_{I\in B}\mathbf{E}\left[\left(r_i^t(I,a) - \tilde{r}_i^t(I,a)\right)^2\right] \tag{110}$$

By Equation 87, $|r_i^t(I,a)| \leq \Delta_{u,i}\pi_{-i}^{\sigma^t}(I)$. From Equation 91, $|\tilde{r}_i^t(I,a)| \leq \Delta_{u,i}^t(I)$. Thus,

$$\mathbf{E}\left[(r_i^t(I,a) - \tilde{r}_i^t(I,a))^2\right] \leq \mathbf{E}\left[(r_i^t(I,a))^2 + (\tilde{r}_i^t(I,a))^2\right] \tag{111}$$

$$\leq \Delta_{u,i}^2\pi_{-i}^{\sigma^t}(I)^2 + \Delta_{u,i}^t(I)^2 \tag{112}$$

Note that for all $B \in \mathcal{B}$, by Lemma 16:

$$\sum_{I\in B}\Delta_{u,i}^2\pi_{-i}^{\sigma^t}(I)^2 \leq \sum_{I\in B}\Delta_{u,i}^2\pi_{-i}^{\sigma^t}(I) \leq \Delta_{u,i}^2\sum_{I\in B}\pi_{-i}^{\sigma^t}(I) \leq \Delta_{u,i}^2 \tag{113}$$

Along with Equation 92, and the fact that $\delta \leq 1$ this means,

$$\sum_{I\in B}\mathbf{E}\left[(r_i^t(I,a) - \tilde{r}_i^t(I,a))^2\right] \leq \Delta_{u,i}^2 + \frac{\Delta_{u,i}^2}{\delta^2} \tag{114}$$

$$\leq 2\frac{\Delta_{u,i}^2}{\delta^2} \tag{115}$$

Returning to Equation 110,

$$\mathbf{E}\left[\left(\sum_{I\in\mathcal{I}_i}(R_i^T(I) - \tilde{R}_i^T(I))\right)^2\right] \leq \frac{|\mathcal{I}_i|}{T^2}\sum_{t=1}^T\sum_{B\in\mathcal{B}_i}\sum_{a\in A(B)} 2\frac{\Delta_{u,i}^2}{\delta^2} \tag{116}$$

$$\leq \frac{2|\mathcal{I}_i|\Delta_{u,i}^2}{\delta^2 T}\sum_{B\in\mathcal{I}_i}|A(B)| \tag{117}$$

Thus by Lemma 2, with probability at least $1-p$,

$$R_i^T \leq \frac{\sqrt{2|\mathcal{I}_i||\mathcal{B}_i||A_i|}\Delta_{u,i}}{\delta\sqrt{pT}} + \frac{\Delta_{u,i}M\sqrt{|A_i|}}{\delta\sqrt{T}} \tag{118}$$

Since $M \geq \sqrt{|I_i||B_i|}$,

$$R_i^T \leq \left(1+\frac{\sqrt{2}}{\sqrt{p}}\right)\left(\frac{1}{\delta}\right)\frac{\Delta_{u,i}M\sqrt{|A_i|}}{\sqrt{T}} \tag{119}$$

■

A.6 Specific MCCFR Variants

We can now apply Theorem 7 to prove a regret bound for outcome-sampling and external-sampling.

A.6.1 Outcome-Sampling

Theorem 8 *For any $p \in (0,1]$, when using outcome-sampling MCCFR where $\forall z \in Z$ either $\pi_{-i}^\sigma(z) = 0$ or $q(z) \geq \delta > 0$ at every timestep, with probability $1-p$, average overall regret is bounded by*

$$R_i^T \leq \left(1+\frac{\sqrt{2}}{\sqrt{p}}\right)\left(\frac{1}{\delta}\right)\frac{\Delta_{u,i}M_i\sqrt{|A_i|}}{\sqrt{T}} \tag{120}$$

Proof: We simply need to show that,

$$\sum_{I\in B}\left(\sum_{z\in Q\cap Z_I}\frac{\pi^\sigma(z[I],z)\pi_{-i}^\sigma(z[I])}{q(z)}\right)^2 \leq \frac{1}{\delta^2}. \tag{121}$$

Note that for all $Q \in \mathcal{Q}$, $|Q| = 1$. Also note that for any $B \in \mathcal{B}_i$ there is at most one $I \in B$ such that $Q \cap Z_I \neq \emptyset$. This is because all the information sets in $Q \cap Z_I$ all have player i's action sequence of a different length, while all information sets in B have player i's action sequence being the same length. Therefore, only a single term of the inner sum is ever non-zero.

Now by our assumption, for all I and $z \in Z_I$ where $\pi_{-i}^\sigma(z) > 0$,

$$\frac{\pi^\sigma(z[I],z)\pi_{-i}^\sigma(z[I])}{q(z)} \leq \frac{1}{\delta} \tag{122}$$

as all the terms of the numerator are less than 1. So the one non-zero term is bounded by $1/\delta$ and so the overall sum of squares must be bounded by $1/\delta^2$. ■

A.6.2 External-Sampling

Theorem 9 *For any $p \in (0,1]$, when using external-sampling MCCFR, with probability at least $1-p$, average overall regret is bounded by*

$$R_i^T \leq \left(1+\frac{\sqrt{2}}{\sqrt{p}}\right)\frac{\Delta_{u,i}M_i\sqrt{|A_i|}}{\sqrt{T}}. \tag{123}$$

Proof: We will simply show that,

$$\sum_{I\in B}\left(\sum_{z\in Q\cap Z_I}\frac{\pi^\sigma(z[I],z)\pi_{-i}^\sigma(z[I])}{q(z)}\right)^2 \leq 1 \tag{124}$$

Since $q(z) = \pi_{-i}^\sigma(z)$, we need to show,

$$\sum_{I\in B}\left(\sum_{z\in Q\cap Z_I}\pi_i^\sigma(z[I],z)\right)^2 \leq 1 \tag{125}$$

Let $\hat{\sigma}^t$ be a deterministic strategy profile sampled from σ^t where Q is the set of histories consistent with $\hat{\sigma}^t_{-i}$. So $Q \cap Z_I \neq \emptyset$ if and only if I is reachable with $\hat{\sigma}^t_{-i}$. By Lemma 10, for all $B \in \mathcal{B}_i$ there is only one $I \in B$ that is reachable; name it I^*. Moreover, there is a unique history in I^* that is a prefix of all $z \in Q \cap Z_{I^*}$; name it h^*. So for all $z \in Q \cap Z_{I^*}$, $z[I^*] = h*$. This is because $\hat{\sigma}^t_t{-i}$ uniquely specifies the actions for all but player i and B uniquely specifies the actions for player i prior to reaching I^*.

Define ρ to be a strategy for all players (including chance) where $\rho_{j \neq i} = \hat{\sigma}_j$ but $\rho_i = \sigma_i$. Consider a $z \in Q \cap Z_I$. z must be reachable by $\hat{\sigma}_{-i}$, so $\pi^\rho_{-i}(z) = 1$. So

$$\sum_{z \in Q \cap Z_{I^*}} \pi^\sigma_i(z[I^*], z) = \sum_{z \in Q \cap Z_{I^*}} \pi^\rho_i(h^*, z) \tag{126}$$

$$= \sum_{z \in Q \cap Z_{I^*}} \pi^\rho(h^*, z) \tag{127}$$

$$\leq \sum_{z \in Z_{I^*}} \pi^\rho(h^*, z) \leq 1 \tag{128}$$

So,

$$\sum_{I \in B} \left(\sum_{z \in Q \cap Z_I} \pi^\sigma_i(z[I], z) \right)^2 \leq 1 \tag{129}$$

■

A.7 Vanilla CFR: A Tighter Bound

In the final proof we use some of the same ideas of the previous proofs to tighten the original bound of vanilla CFR, so the bound depends on M_i rather than $|\mathcal{I}_i|$ as with the MCCFR variants.

Theorem 10 *When using vanilla CFR for player i,* $R^T_i \leq \Delta_{u,i} M_i \sqrt{|A_i|}/\sqrt{T}$.

Proof: Define $\Delta^t_{u,i}(I) = \sigma^t_{-i}(I)\Delta_{u,i}(I)$. Using Theorem 6,

$$(R^{T,+}_i(I))^2 \leq \frac{|A(I)|}{T^2} \sum_{t=1}^{T} (\Delta^t_{u,i}(I))^2 \tag{130}$$

$$R^{T,+}_i(I) \leq \frac{\sqrt{|A(I)|}\Delta_{u,i}(I)}{T} \sqrt{\sum_{t=1}^{T} (\sigma^t_{-i}(I))^2}. \tag{131}$$

By summing over all information sets of I, we get:

$$R^{T,+}_i \leq \frac{1}{T} \sum_{I \in \mathcal{I}_i} \sqrt{|A(I)|}\Delta_{u,i}(I) \sqrt{\sum_{t=1}^{T} (\sigma^t_{-i}(I))^2} \tag{132}$$

$$\leq \frac{\sqrt{|A_i|}\Delta_{u,i}}{T} \sum_{I \in \mathcal{I}_i} \sqrt{\sum_{t=1}^{T} (\sigma^t_{-i}(I))^2} \tag{133}$$

$$\leq \frac{\sqrt{|A_i|}\Delta_{u,i}}{T} \sum_{B \in \mathcal{B}_i} \sum_{I \in B} \sqrt{\sum_{t=1}^{T} (\sigma^t_{-i}(I))^2}. \tag{134}$$

For each action sequence $B \in \mathcal{B}_i$:

$$\sum_{I \in B} \sigma^t_{-i}(I) \leq 1 \tag{135}$$

$$\sum_{t=1}^{T} \sum_{I \in B} \sigma^t_{-i}(I) \leq T \tag{136}$$

Therefore, by Lemma 6:

$$\sum_{I \in B} \sum_{t=1}^{T} \sqrt{\sigma_{-i}^{t}(I)} \leq \sqrt{|B|T} \tag{137}$$

Summing over all $B \in \mathcal{B}_i$ yields:

$$R_i^{T,+} \leq \frac{\sqrt{|A_i|}\Delta_{u,i}}{T} \sum_{B \in \mathcal{B}_i} \sqrt{|B||T|}. \tag{138}$$

■

In practice, this makes the bound on vanilla counterfactual regret as tight as the sampling bounds. The distinctive difference is the amount of computation required per iteration.