Supplemental Material for *Monte Carlo Sampling for Regret Minimization in Extensive Games*

Marc Lanctot
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada T6G 2E8
lanctot@ualberta.ca

Kevin Waugh
School of Computer Science
Carnegie Mellon University
Pittsburgh PA 15213-3891
waugh@cs.cmu.edu

Martin Zinkevich
Yahoo! Research
Santa Clara, CA, USA 95054
maz@yahoo-inc.com

Michael Bowling
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada T6G 2E8
bowling@cs.ualberta.ca

1 Introduction

The supplementary material presented here first presents a detailed description of the MCCFR algorithm. We then give proofs to Theorems 3, 4, and 5 from the submission *Monte Carlo Sampling for Regret Minimization in Extensive Games*. We begin with some preliminaries, then prove a general result about all members of the MCCFR family of algorithms (Theorem 18 in Section 6). We then use that result to prove bounds for the MCCFR variants (Theorems 19 and 20 in Section 7). We finally prove the tightened bound for vanilla CFR (Theorem 21 in Section 8).

2 MCCFR Algorithm

The MCCFR algorithm is presented in detail in Algorithm 1.

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*.

2.1 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

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

our sampled history) then

$$\tilde{r}(I, a) = \tilde{v}_i(\sigma^t_{(I \to a)}, I) - \tilde{v}_i(\sigma^t, I) \tag{1}$$

$$= \frac{u_i(z)\pi^\sigma_{-i}(z[I])\pi^\sigma(z[I]a, z)}{\pi^{\sigma'}(z)} - \frac{u_i(z)\pi^\sigma_{-i}(z[I])\pi^\sigma(z[I], z)}{\pi^{\sigma'}(z)} \tag{2}$$

$$= \frac{u_i(z)\pi^\sigma_{-i}(z[I])}{\pi^{\sigma'}(z)} (\pi^\sigma(z[I]a, z) - \pi^\sigma(z[I], z)) \tag{3}$$

$$= W \cdot (\pi^\sigma(z[I]a, z) - \pi^\sigma(z[I], z)) \tag{4}$$

where

$$W = \frac{u_i(z)\pi^\sigma_{-i}(z[I])}{\pi^{\sigma'}(z)} \tag{5}$$

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

$$\tilde{r}(I, a) = 0 - \tilde{v}_i(\sigma^t, I) \tag{6}$$

$$= -W \cdot \pi^\sigma(z[I], z) \tag{7}$$

2.2 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:

$$\tilde{r}(I,a) = \sum_{z \in Q \cap Z_I} \left(\tilde{v}_i(\sigma^t_{(I \to a)}, I) - \tilde{v}_i(\sigma^t, I) \right) \tag{8}$$

$$= \sum_{z \in Q \cap Z_I} \frac{u_i(z)\pi^\sigma_{-i}(z[I])}{q(z)} \left(\pi^\sigma(z[I]a, z) - \pi^\sigma(z[I], z) \right) \tag{9}$$

$$= \sum_{z \in Q \cap Z_I} \frac{u_i(z)\pi^\sigma_{-i}(z[I])\pi^\sigma_{-i}(z[I]a, z)}{q(z)} \left(\frac{\pi^\sigma(z[I]a, z) - \pi^\sigma(z[I], z)}{\pi^\sigma_{-i}(z[I]a, z)} \right) \tag{10}$$

$$= \sum_{z \in Q \cap Z_I} \frac{u_i(z)\pi^\sigma_{-i}(z)}{q(z)} \left(\pi^\sigma_i(z[I]a, z) - \pi^\sigma_i(z[I], z) \right) \tag{11}$$

$$= \sum_{z \in Q \cap Z_I} u_i(z)\pi^\sigma_i(z[I]a, z) \left(1 - \sigma(a|I) \right) \tag{12}$$

$$= \left(1 - \sigma(a|I) \right) \sum_{z \in Q \cap Z_I} u_i(z)\pi^\sigma_i(z[I]a, z) \tag{13}$$

$$= \frac{\sigma(a|I)}{\sigma(a|I)} \left(1 - \sigma(a|I) \right) \sum_{z \in Q \cap Z_I} u_i(z)\pi^\sigma_i(z[I]a, z) \tag{14}$$

$$= \left(\frac{1}{\sigma(a|I)} - 1 \right) \sum_{z \in Q \cap Z_I} u_i(z)\pi^\sigma_i(z[I], z) \tag{15}$$

$$= \frac{1}{\sigma(a|I)} \sum_{z \in Q \cap Z_I} u_i(z)\pi^\sigma_i(z[I], z) - \sum_{z \in Q \cap Z_I} u_i(z)\pi^\sigma_i(z[I], z) \tag{16}$$

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.

3 Preliminaries

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

Lemma 1 *For any random variable X:*

$$\Pr[|X| \geq k\sqrt{\mathbf{E}[X^2]}] \leq \frac{1}{k^2}. \tag{17}$$

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

$$\Pr[Y \geq j\mathbf{E}[Y]] \leq \frac{1}{j}. \tag{18}$$

By setting $Y = X^2$:

$$\Pr[X^2 \geq j\mathbf{E}[X^2]] \leq \frac{1}{j} \tag{19}$$

$$\Pr[|X| \geq \sqrt{j\mathbf{E}[X^2]}] \leq \frac{1}{j}. \tag{20}$$

Replacing $k = \sqrt{j}$:

$$\Pr[|X| \geq k\sqrt{\mathbf{E}[X^2]}] \leq \frac{1}{j^2}. \tag{21}$$

■

Lemma 2 *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:

$$(a_i)^2 + (a_j)^2 \leq (a_i)^2 + 2a_i a_j + (a_j)^2 \tag{22}$$
$$\leq (a_i + a_j)^2. \tag{23}$$

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:

$$0 = f'(x) \tag{24}$$
$$f'(x) = -2(A - x) + 2x \tag{25}$$
$$2A = 4x \tag{26}$$
$$\frac{A}{2} = x \tag{27}$$

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 3 *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:

$$f'(x) = 0 \tag{28}$$
$$f'(x) = \frac{0.5(n-1)}{\sqrt{x(n-1)}} - \frac{0.5}{\sqrt{S - x}} \tag{29}$$
$$\frac{0.5\sqrt{n-1}}{\sqrt{x}} = \frac{0.5}{\sqrt{S - x}} \tag{30}$$
$$\frac{x}{n-1} = S - x \tag{31}$$
$$x\left(1 + \frac{1}{n-1}\right) = S \tag{32}$$
$$x\left(\frac{n-1+1}{n-1}\right) = S \tag{33}$$
$$x = \frac{S(n-1)}{n} \tag{34}$$

Therefore, substituting the three critical points yields:

$$f(0) = \sqrt{S} \tag{35}$$

$$f(S) = \sqrt{S(n-1)} \tag{36}$$

$$f\left(\frac{S(n-1)}{n}\right) = \sqrt{\frac{S(n-1)(n-1)}{n}} + \sqrt{S - \frac{S(n-1)}{n}} \tag{37}$$

$$= (n-1)\sqrt{\frac{S}{n}} + \sqrt{\frac{S}{n}} \tag{38}$$

$$= \sqrt{Sn} \tag{39}$$

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

Lemma 4 *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 3. ∎

Lemma 5 *Given nonnegative reals $a_{i,j}$ in $[0,1]$, where $\sum_{i=1}^m \sum_{j=1}^n a_{m,n} = S$, then:*

$$\sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} (a_{m,n})^2} \leq \sqrt{mS}. \tag{40}$$

4 Blackwell's Approachability Theorem

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

Lemma 6 *For all real a, define $a^+ = \max(a, 0)$. For all a, b, it is the case that*

$$\left((a+b)^+\right)^2 \leq (a^+)^2 + 2(a^+)b + b^2 \tag{41}$$

Proof: We prove this by enumerating the possibilities:

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

$$\left((a+b)^+\right)^2 \leq (b^+)^2 \tag{42}$$

$$\leq b^2, \tag{43}$$

and:

$$(a^+)^2 + 2(a^+)b + b^2 = b^2. \tag{44}$$

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

$$\left((a+b)^+\right)^2 = (a+b)^2. \tag{45}$$

Also:

$$(a^+)^2 + 2(a^+)b + b^2 = a^2 + 2ab + b^2 \tag{46}$$

$$= (a+b)^2 \tag{47}$$

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

$$\left((a+b)^+\right)^2 = 0. \tag{48}$$

Also:

$$(a^+)^2 + 2(a^+)b + b^2 = a^2 + 2ab + b^2 \tag{49}$$

$$= (a+b)^2 \tag{50}$$

$$\geq 0 \tag{51}$$

■

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

$$\sigma_{T+1}(a) = \begin{cases} \frac{R^+_T(a)}{R^+_{\sum,T}} \text{ if } R^+_{\sum,T} > 0 \\ \frac{1}{|A|} \quad \text{otherwise} \end{cases} \tag{52}$$

Lemma 7 *If regret matching is used, then:*

$$\sum_{a\in A} R^+_T(a) r_{T+1}(a) \leq 0 \tag{53}$$

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

$$\begin{aligned}
\sum_{a\in A} R^+_T(a) r_{T+1}(a) &= \sum_{a\in A} R^+_T(a)(u_{T+1}(a) - u_{T+1}(\sigma_t)) && (54)\\
&= \left(\sum_{a\in A} R^+_T(a) u_{T+1}(a)\right) - \left(u_{T+1}(\sigma_t) \sum_{a\in A} R^+_T(a)\right) && (55)\\
&= \left(\sum_{a\in A} R^+_T(a) u_{T+1}(a)\right) - \left(\sum_{a'\in A} \sigma_{T+1}(a') u_{T+1}(a')\right) R^+_{\sum,T} && (56)\\
&= \left(\sum_{a\in A} R^+_T(a) u_{T+1}(a)\right) - \left(\sum_{a'\in A} \frac{R^+_T(a')}{R^+_{\sum,T}} u_{T+1}(a')\right) R^+_{\sum,T} && (57)\\
&= \left(\sum_{a\in A} R^+_T(a) u_{T+1}(a)\right) - \left(\sum_{a'\in A} R^+_T(a') u_{T+1}(a')\right) && (58)\\
&= 0 && (59)
\end{aligned}$$

■

Theorem 8 *Define Δ_t to be $\max_{a,a'\in A}(u_t(a) - u_t(a'))$. Then regret matching yields:*

$$\sum_{a\in A} (R^+_T(a))^2 \leq \frac{1}{T^2} \sum_{t=1}^{T} |A| (\Delta_t)^2. \tag{60}$$

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 6:

$$(R^+_T(a))^2 \leq (\frac{(T-1)R^+_{T-1}(a)}{T})^2 + 2\frac{T-1}{T^2} R^+_{T-1}(a) r_T(a) + (\frac{r_T(a)}{T})^2 \tag{61}$$

Summing yields:

$$\sum_{a\in A} (R^+_T(a))^2 \leq \sum_{a\in A} \left(\left(\frac{T-1}{T}\right)^2 (R^+_{T-1}(a))^2 + 2\frac{T-1}{T^2} R^+_{T-1}(a) r_T(a) + \frac{1}{T^2}(r_T(a))^2\right) \tag{62}$$

By Lemma 7, $\sum_{a\in A} R^+_{T-1}(a) r_T(a) = 0$, so:

$$\sum_{a\in A} (R^+_T(a))^2 \leq \left(\left(\frac{T-1}{T}\right)^2 \sum_{a\in A} (R^+_{T-1}(a))^2\right) + \left(\frac{1}{T^2} \sum_{a\in A} (r_T(a))^2\right) \tag{63}$$

By induction:

$$\sum_{a\in A} (R^+_{T-1}(a))^2 \leq \frac{1}{(T-1)^2} \sum_{t=1}^{T-1} |A| (\Delta_t)^2. \tag{64}$$

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

$$\sum_{a \in A} (R_T^+(a))^2 \leq \frac{1}{T^2} \left(\sum_{t=1}^{T-1} |A| (\Delta_t)^2 \right) + |A| (\Delta_T)^2. \tag{65}$$

■

5 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 $t \in \{0 \dots |h| - 1\}$, $\sigma(h(t)) = h_{t+1}$. 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 $t \in \{|h'| \dots |h| - 1\}$, $\sigma(h(t)) = h_{t+1}$. 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 $\boldsymbol{\vec{a}}$** 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 $h \in H$,*

$$\Pr_{\hat{\sigma}_j \in \rho_j}[\forall (I,a) \in X_j(h), \hat{\sigma}_j(I) = a] = \pi^{\sigma_j}(h). \tag{66}$$

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

$$\rho(\hat{\sigma}_j) = \prod_{I \in \mathcal{I}_j} \sigma_j(I)(\hat{\sigma}_j(I)). \tag{67}$$

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:*

$$\Pr_{\hat{\sigma}_i \in \rho}[\text{Reach}_i^{\hat{\sigma}_i}(h)] = \pi_i^{\sigma_i}(h). \tag{68}$$

Lemma 13 *Given a history $h \in H$, $\text{Reach}_i^{\hat{\sigma}_i}(h)$ 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 $\text{Reach}_i^{\hat{\sigma}_i}(h)$, 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 14 *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 15 *Given a set of strategies $\hat{\sigma}_S = \{\hat{\sigma}_i\}_{i \in S}$, if for all $i \in S$, $\text{Reach}_i^{\hat{\sigma}_i}(h)$, then $\text{Reach}_S^{\hat{\sigma}_S}(h)$.*

Proof: By Corollary 14, 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 16 *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,

$$\rho(\hat{\sigma}_{-i}) = \prod_{j \in N' \backslash i} \rho_j(\hat{\sigma}_j). \tag{69}$$

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:

$$\Pr_{\hat{\sigma}_{-i} \in \rho}[\forall j \in N' \backslash i, \text{Reach}_j^{\hat{\sigma}_j}(h)] = \prod_{j \in N' \backslash i} \Pr_{\hat{\sigma}_j \in \rho_j}[\text{Reach}_j^{\hat{\sigma}_j}(h)] \tag{70}$$

$$= \prod_{j \in N' \backslash i} \pi^{\hat{\sigma}_j}(h) \tag{71}$$

$$= \pi^{\hat{\sigma}_{-i}}(h) \tag{72}$$

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

Lemma 17 *For any strategy profile σ_{-i}, for any $\vec{a}$:*

$$\sum_{I \in \mathcal{I}_i(\vec{a})} \pi_{-i}^{\sigma_{-i}}(I) \leq 1 \tag{73}$$

Proof: This follows directly from Lemma 16 and Lemma 10. ∎

6 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 18 *For any $p \in (0, 1]$, when using any algorithm in the MCCFR family such that for all $Q \in \mathcal{Q}$ and $B \in \mathcal{B}$,*

$$\sum_{I \in B} \left(\sum_{z \in Q \cap Z_I} \frac{\pi^{\sigma}(z[I], z) \pi_{-i}^{\sigma}(z[I])}{q(z)} \right)^2 \leq \frac{1}{\delta^2} \tag{74}$$

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

$$R_i^T \leq \left(1 + \frac{2}{\sqrt{p}}\right) \left(\frac{1}{\delta}\right) \frac{\Delta_{u,i} M_i \sqrt{|A_i|}}{\sqrt{T}}. \tag{75}$$

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,

$$r_i^t(I, a) = \left(v_i(\sigma^t_{(I \to a)}, I) - v_i(\sigma^t, I)\right) \tag{76}$$

$$\tilde{r}_i^t(I, a) = \left(\tilde{v}_i(\sigma^t_{(I \to a)}, I) - \tilde{v}_i(\sigma^t, I)\right) \tag{77}$$

$$R_i^T(I) = \frac{1}{T} \max_{a \in A(I)} \sum_{t=1}^{T} r_i^t(I, a) \tag{78}$$

$$\tilde{R}_i^T(I) = \frac{1}{T} \max_{a \in A(I)} \sum_{t=1}^{T} \tilde{r}_i^t(I, a) \tag{79}$$

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,

$$\left(\tilde{v}_i(\sigma^t_{(I \to a)}, I) - \tilde{v}_i(\sigma^t, I)\right) \leq \Delta_{u,i}^t(I) \equiv \Delta_{u,i} \sum_{z \in Q_t \cap Z_I} \frac{\pi^{\sigma}(z[I], z) \pi_{-i}^{\sigma}(z[I])}{q(z)} \tag{80}$$

so by our assumption,

$$\sum_{I \in B} \Delta_{u,i}^t(I)^2 \leq \frac{\Delta_{u,i}^2}{\delta^2} \tag{81}$$

So we can apply Theorem 8, to get,

$$\tilde{R}_i^T(I) \leq \frac{\sqrt{|A(I)| \sum_{t=1}^{T} (\Delta_{u,i}^t(I))^2}}{T} \tag{82}$$

Using Lemma 4,

$$\sum_{I \in B} \tilde{R}_i^T(I) \leq \frac{\sqrt{|B||A(B)| \sum_{I \in B} \sum_{t=1}^{T} (\Delta_{u,i}^t(I))^2}}{T} \tag{83}$$

$$\leq \frac{\sqrt{|B||A(B)| \sum_{t=1}^{T} \sum_{I \in B} (\Delta_{u,i}^t(I))^2}}{T} \tag{84}$$

$$\leq \frac{\sqrt{|B||A(B)| \sum_{t=1}^{T} \Delta_{u,i}^2 / \delta^2}}{T} \tag{85}$$

$$\leq \frac{\Delta_{u,i} \sqrt{|B||A(B)|}}{\delta \sqrt{T}} \tag{86}$$

The average overall sampled regret then can be bounded by,

$$\tilde{R}_i^T \leq \sum_{B \in \mathcal{B}_i} \sum_{I \in B} \tilde{R}_i^T(I) \tag{87}$$

$$\leq \sum_{B \in \mathcal{B}_i} \frac{\Delta_{u,i} \sqrt{|B||A(B)|}}{\delta \sqrt{T}} \tag{88}$$

$$\leq \frac{\Delta_{u,i} \sqrt{|A_i|} \sum_{B \in \mathcal{B}_i} \sqrt{|B|}}{\delta \sqrt{T}} \tag{89}$$

$$\leq \frac{\Delta_{u,i} M_i \sqrt{|A_i|}}{\delta \sqrt{T}} \tag{90}$$

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 $\sum_{I \in \mathcal{I}_i} R_i^T(I)$ and $\sum_{I \in \mathcal{I}_i} \tilde{R}_i^T(I)$ in order to prove that they are close, and then use Lemma 1 to bound the absolute value. We begin by focusing on the similarity of the counterfactual regret ($R_i^T(I)$ and $\tilde{R}_i^T(I)$) in every node, by focusing on the similarity of the counterfactual regret of a particular action at a particular time ($r_i^t(I,a)$ and $\tilde{r}_i^t(I,a)$). By the Lemma from the main paper, we know that $\mathbf{E}[r_i^t(I,a) - \tilde{r}_i^t(I,a)] = 0$.

From Lemma 4 we have,

$$\mathbf{E}\left[\left(\sum_{I \in \mathcal{I}_i} (R_i^T(I) - \tilde{R}_i^T(I))\right)^2\right] \leq |\mathcal{I}_i| \sum_{I \in \mathcal{I}_i} \mathbf{E}\left[(R_i^T(I) - \tilde{R}_i^T(I))^2\right] \tag{91}$$

So,

$$(R_i^T(I) - \tilde{R}_i^T(I))^2 = \left(\frac{1}{T} \max_{a \in A(I)} \sum_{t=1}^{T} r_i^t(I,a) - \frac{1}{T} \max_{a \in A(I)} \sum_{t=1}^{T} \tilde{r}_i^t(I,a)\right)^2 \tag{92}$$

$$(R_i^T(I) - \tilde{R}_i^T(I))^2 \leq \frac{1}{T^2} \left(\max_{a \in A(I)} \left(\sum_{t=1}^{T} r_i^t(I,a) - \sum_{t=1}^{T} \tilde{r}_i^t(I,a)\right)\right)^2 \tag{93}$$

$$(R_i^T(I) - \tilde{R}_i^T(I))^2 \leq \frac{1}{T^2} \left(\max_{a \in A(I)} \left(\sum_{t=1}^{T} \left|r_i^t(I,a) - \tilde{r}_i^t(I,a)\right|\right)\right)^2 \tag{94}$$

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

$$(R_i^T(I) - \tilde{R}_i^T(I))^2 \leq \frac{1}{T^2} \max_{a \in A(I)} \left(\sum_{t=1}^{T} r_i^t(I,a) - \sum_{t=1}^{T} \tilde{r}_i^t(I,a) \right)^2 \tag{95}$$

$$(R_i^T(I) - \tilde{R}_i^T(I))^2 \leq \frac{1}{T^2} \sum_{a \in A(I)} \left(\sum_{t=1}^{T} r_i^t(I,a) - \sum_{t=1}^{T} \tilde{r}_i^t(I,a) \right)^2 \tag{96}$$

$$\mathbf{E}[(R_i^T(I) - \tilde{R}_i^T(I))^2] \leq \frac{1}{T^2} \sum_{a \in A(I)} \sum_{t=1}^{T} \mathbf{E}[\left(r_i^t(I,a) - \tilde{r}_i^t(I,a)\right)^2] \tag{97}$$

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 91:

$$\mathbf{E}\left[(\sum_{I \in \mathcal{I}_i} (R_i^T(I) - \tilde{R}_i^T(I)))^2 \right] \leq \frac{|\mathcal{I}_i|}{T^2} \sum_{I \in \mathcal{I}_i} \sum_{a \in A(I)} \sum_{t=1}^{T} \mathbf{E}\left[\left(r_i^t(I,a) - \tilde{r}_i^t(I,a)\right)^2 \right] \tag{98}$$

$$\leq \frac{|\mathcal{I}_i|}{T^2} \sum_{t=1}^{T} \sum_{B \in \mathcal{B}_i} \sum_{a \in A(B)} \sum_{I \in B} \mathbf{E}\left[\left(r_i^t(I,a) - \tilde{r}_i^t(I,a)\right)^2 \right] \tag{99}$$

By Equation 76, $|r_i^t(I,a)| \leq \Delta_{u,i} \pi_{-i}^{\sigma^t}(I)$. From Equation 80, $|\tilde{r}_i^t(I,a)| \leq \Delta_{u,i}^t(I)$. Thus,

$$\mathbf{E}\left[(r_i^t(I,a) - \tilde{r}_i^t(I,a))^2 \right] \leq \mathbf{E}\left[(r_i^t(I,a))^2 + (\tilde{r}_i^t(I,a))^2 \right] \tag{100}$$

$$\leq \Delta_{u,i}^2 \pi_{-i}^{\sigma^t}(I)^2 + \Delta_{u,i}^t(I)^2 \tag{101}$$

Note that for all $B \in \mathcal{B}$, by Lemma 17:

$$\sum_{I \in B} \Delta_{u,i}^2 \pi_{-i}^{\sigma^t}(I)^2 \leq \sum_{I \in B} \Delta_{u,i}^2 \pi_{-i}^{\sigma^t}(I) \leq \Delta_{u,i}^2 \sum_{I \in B} \pi_{-i}^{\sigma^t}(I) \leq \Delta_{u,i}^2 \tag{102}$$

Along with Equation 81, and the fact that $\delta \leq 1$ this means,

$$\sum_{I \in B} \mathbf{E}\left[(r_i^t(I,a) - \tilde{r}_i^t(I,a))^2 \right] \leq \Delta_{u,i}^2 + \frac{\Delta_{u,i}^2}{\delta^2} \tag{103}$$

$$\leq 2 \frac{\Delta_{u,i}^2}{\delta^2} \tag{104}$$

Returning to Equation 99,

$$\mathbf{E}\left[\left(\sum_{I \in \mathcal{I}_i} (R_i^T(I) - \tilde{R}_i^T(I)) \right)^2 \right] \leq \frac{|\mathcal{I}_i|}{T^2} \sum_{t=1}^{T} \sum_{B \in \mathcal{B}_i} \sum_{a \in A(B)} 2 \frac{\Delta_{u,i}^2}{\delta^2} \tag{105}$$

$$\leq \frac{2|\mathcal{I}_i| \Delta_{u,i}^2}{\delta^2 T} \sum_{B \in \mathcal{I}_i} |A(B)| \tag{106}$$

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

$$R_i^T \leq \frac{\sqrt{2|\mathcal{I}_i||\mathcal{B}_i||A_i|}\Delta_{u,i}}{\delta\sqrt{pT}} + \frac{\Delta_{u,i} M \sqrt{|A_i|}}{\delta\sqrt{T}} \tag{107}$$

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

$$R_i^T \leq \left(1 + \frac{\sqrt{2}}{\sqrt{p}}\right) \left(\frac{1}{\delta}\right) \frac{\Delta_{u,i} M \sqrt{|A_i|}}{\sqrt{T}} \tag{108}$$

■

7 Specific MCCFR Variants

We can now apply Theorem 18 to prove a regret bound for outcome-sampling and external-sampling.

7.1 Outcome-Sampling

Theorem 19 *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*

$$R_i^T \leq \left(1 + \frac{\sqrt{2}}{\sqrt{p}}\right)\left(\frac{1}{\delta}\right)\frac{\Delta_{u,i}M_i\sqrt{|A_i|}}{\sqrt{T}} \tag{109}$$

Proof: We simply need to show that,

$$\sum_{I \in B}\left(\sum_{z \in Q \cap Z_I}\frac{\pi^{\sigma}(z[I], z)\pi^{\sigma}_{-i}(z[I])}{q(z)}\right)^2 \leq \frac{1}{\delta^2}. \tag{110}$$

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 $\pi^{\sigma}_{-i}(z) > 0$,

$$\frac{\pi^{\sigma}(z[I], z)\pi^{\sigma}_{-i}(z[I])}{q(z)} \leq \frac{1}{\delta} \tag{111}$$

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 $1/\delta^2$. ∎

7.2 External-Sampling

Theorem 20 *For any $p \in (0, 1]$, when using external-sampling MCCFR, with probability at least $1 - p$, average overall regret is bounded by*

$$R_i^T \leq \left(1 + \frac{\sqrt{2}}{\sqrt{p}}\right)\frac{\Delta_{u,i}M_i\sqrt{|A_i|}}{\sqrt{T}}. \tag{112}$$

Proof: We will simply show that,

$$\sum_{I \in B}\left(\sum_{z \in Q \cap Z_I}\frac{\pi^{\sigma}(z[I], z)\pi^{\sigma}_{-i}(z[I])}{q(z)}\right)^2 \leq 1 \tag{113}$$

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

$$\sum_{I \in B}\left(\sum_{z \in Q \cap Z_I}\pi^{\sigma}_i(z[I], z)\right)^2 \leq 1 \tag{114}$$

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

$$\sum_{z\in Q\cap Z_{I^*}} \pi^{\sigma}_i(z[I^*], z) = \sum_{z\in Q\cap Z_{I^*}} \pi^{\rho}_i(h^*, z) \tag{115}$$

$$= \sum_{z\in Q\cap Z_{I^*}} \pi^{\rho}(h^*, z) \tag{116}$$

$$\leq \sum_{z\in Z_{I^*}} \pi^{\rho}(h^*, z) \leq 1 \tag{117}$$

So,

$$\sum_{I\in B} \left(\sum_{z\in Q\cap Z_I} \pi^{\sigma}_i(z[I], z) \right)^2 \leq 1 \tag{118}$$

■

8 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 21 *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 8,

$$(R^{T,+}_i(I))^2 \leq \frac{|A(I)|}{T^2} \sum_{t=1}^{T} (\Delta^t_{u,i}(I))^2 \tag{119}$$

$$R^{T,+}_i(I) \leq \frac{\sqrt{|A(I)|}\Delta_{u,i}(I)}{T} \sqrt{\sum_{t=1}^{T} (\sigma^t_{-i}(I))^2}. \tag{120}$$

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

$$R^{T,+}_i \leq \frac{1}{T} \sum_{I\in\mathcal{I}_i} \sqrt{|A(I)|}\Delta_{u,i}(I) \sqrt{\sum_{t=1}^{T} (\sigma^t_{-i}(I))^2} \tag{121}$$

$$\leq \frac{\sqrt{|A_i|}\Delta_{u,i}}{T} \sum_{I\in\mathcal{I}_i} \sqrt{\sum_{t=1}^{T} (\sigma^t_{-i}(I))^2} \tag{122}$$

$$\leq \frac{\sqrt{|A_i|}\Delta_{u,i}}{T} \sum_{B\in\mathcal{B}_i} \sum_{I\in B} \sqrt{\sum_{t=1}^{T} (\sigma^t_{-i}(I))^2}. \tag{123}$$

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

$$\sum_{I\in B} \sigma^t_{-i}(I) \leq 1 \tag{124}$$

$$\sum_{t=1}^{T} \sum_{I\in B} \sigma^t_{-i}(I) \leq T \tag{125}$$

Therefore, by Lemma 5:

$$\sum_{I\in B} \sum_{t=1}^{T} \sqrt{\sigma^t_{-i}(I)} \leq \sqrt{|B|T} \tag{126}$$

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

$$R_i^{T,+} \leq \frac{\sqrt{|A_i|}\Delta_{u,i}}{T} \sum_{B \in \mathcal{B}_i} \sqrt{|B||T|}. \tag{127}$$

■

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.