Monte Carlo Sampling for Regret Minimization in Extensive Games

Marc Lanctot, Kevin Waugh, and Michael Bowling
University of Alberta
Edmonton, Alberta, Canada
{lanctot|waugh|bowling}@cs.ualberta.ca

Abstract

One efficient method for computing Nash equilibria in large, zero-sum extensive games is counterfactual regret minimization (CFR). In the domain of poker, CFR has proven effective, particularly when using a domain-specific augmentation involving chance outcome sampling. In this paper we introduce MCCFR, a Monte Carlo version of the algorithm capable of applying equivalent updates (in expectation) on sampled histories which has bounded overall regret with high probability. We show empirically that, although MCCFR requires more iterations, its lower cost per iteration results in overall faster convergence, particularly as the game size increases.

1 Introduction

The past few years have seen dramatic algorithmic improvements in finding approximate Nash equilibria, in two-player, zero-sum extensive games [ZJBP08; GHPS07]. *Counterfactual regret minimization* (CFR) is one recent technique that exploits the fact that the time-averaged strategy profile of regret minimizing algorithms converges to a Nash equilibrium; it has been been successfully applied to Poker games which have up to 10^{12} game states. The key insight is the fact that minimizing per-information set counterfactual regret results in minimizing overall regret.

The vanilla form presented by Zinkevich and colleagues requires the entire game tree to be traversed on each iteration. Fortunately, it is possible to avoid a full game-tree traversal. In their accompanying technical report, Zinkevich and colleagues discuss a *poker-specific* CFR variant that samples chance outcomes on each iteration [ZBJP07]. 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. However, their chance-sampling variant and its derived bound are limited to poker-like games.

An additional disadvantage of CFR is that it requires the opponent's policy to be known, which makes it unsuitable for general regret minimization in an extensive game. General regret minimization in extensive games is possible using online convex programming techniques, such as Lagrangian Hedging [Gor07], 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 CFR minimizing algorithms 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 Monte Carlo CFR (MCCFR) as the most extreme form of sampling in this family: only a single playing of the game is sampled on each iteration. We show that under a reasonable sampling strategy, any member of this family minimizes overall regret, and so can be used for equilibrium computation. In addition, the MCCFR algorithm does not need knowledge of opponent probabilities beyond samples of play from the strategy. This makes MCCFR suitable for general regret minimization in online extensive game settings.

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, $I_i \in \mathcal{I}_i$, 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.

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.

A **history**, $h \in H$, is a sequence of actions; a terminal history or playout $z \in Z$ is a history that leads to a leaf in the game tree. Note that the perfect recall assumption implies that every history can be expressed as the empty history or a concatenation of a history and an action choice. A **strategy for player i**, σ_i, is a function that assigns a probability distribution over all actions at each information set belonging to i and by convention the opponent's strategy is denoted σ_{-i}. A **strategy profile** σ is a collection of strategies for each player. A strategy profile is called a **Nash equilibrium** if each player has no incentive to deviate unilaterally. Equiv-

alently, each strategy is not exploitable: any **best response strategy** to σ_{-i} will result in no gain in utility when played against σ_{-i}. An **ϵ-equilibrium** is one where, for each σ_i the best response to σ_{-i} may result in a gain of at most ϵ.

2.1 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. There is a well-known connection between regret, average strategies $\bar{\sigma}$, and the Nash equilibrium solution concept.

Theorem 1 *In a zero-sum game at time T, if both player's average overall regret $R_i^T < \epsilon$ 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. As a result, 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 rate of convergence of the approximation.

Zinkevich and colleagues used the above approach in their counterfactual regret algorithm (CFR) [ZJBP08]. 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. Define **counterfactual value** $v_i(\sigma, I)$ as,

$$v_i(\sigma, I) = \sum_{h \in I, z \in Z} \pi_{-i}^{\sigma}(h) \pi^{\sigma}(h, z) u_i(z). \tag{1}$$

where $\pi_{-i}^{\sigma}(h)$ and $\pi^{\sigma}(h, z)$ are product of probabilities of the opponent's strategy over h and of both player's probabilities from h to z, and $u_i(z)$ is the payoff to i for playout z. The **immediate counterfactual regret** is then,

$$R_{i,\text{imm}}^T(I) = \max_{a \in A(I)} R_{i,\text{imm}}^T(I, a) \tag{2}$$

$$R_{i,\text{imm}}^T(I, a) = \frac{1}{T} \sum_{t=1}^{T} \left(v_i(\sigma_{(I \to a)}^t, I) - v_i(\sigma^t, I) \right) \tag{3}$$

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

Theorem 2 *[ZJBP08, Theorem 3]*
$R_i^T \leq \sum_{I \in \mathcal{I}_i} R_{i,\text{imm}}^{T,+}(I)$

Using Blackwell's algorithm for approachability [Bla56] the positive per-information set immediate counterfactual regrets can be driven to zero by simply normalizing the positive parts of their accumulated values, thus driving average overall regret to zero.

Theorem 3 *[ZJBP08, Theorem 4] Using Blackwell's algorithm to minimize regret with respect to counterfactual value at each information set leads to*

$$R_i^T \leq \Delta_{u,i} |\mathcal{I}_i| \sqrt{|A_i|} / \sqrt{T} \tag{4}$$

where $|A_i|$ is the maximum number of actions i can take at any of their information sets and $\Delta_{u,i} = max_z u_i(z) - min_z u_i(z)$ is the payoff range for player i.

Theorem 3 can be directly turned into an algorithm for computing an approximate Nash equilibrium, which we call *vanilla CFR*. The idea is to traverse the game tree computing counterfactual values using Equation 1. Given a strategy, these values define regret terms for each player for each of their information sets using Equation 3. These regret values accumulate and determine the strategies at the next iteration using Blackwell's formula. Theorem 3 bounds both players' average overall regret (bound decreasing with the number of iterations), which from Theorem 1 means that the average strategy profile $\bar{\sigma}^t$ converges to a Nash Equilibrium.

3 Sample-Based 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 partition of Z. On each iteration we will sample one block of this partition 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 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. The **sampled counterfactual value** when updating block j is:

$$\tilde{v}_i(\sigma, I|j) = \frac{1}{q_j} \sum_{z \in Q_j \cap Z_I} u_i(z) \pi_{-i}^{\sigma}(z[I]) \pi^{\sigma}(z[I], z) \tag{5}$$

Selecting a partition $\mathcal{Q}$ along with the sampling probabilities defines a complete sample-based CFR algorithm. Rather than doing full game tree traversals the algorithm samples one of the blocks of the partition and examines terminal histories in that block only. Note that we get vanilla CFR when $\mathcal{Q} = \{Z\}$ and chance-sampled CFR when $\mathcal{Q}$ splits sets into terminal histories by chance node outcomes.

Lemma 4 *Sampled counterfactual value equals counterfactual value in expectation. Formally,*

$$E_{j \sim q_j} \left[\tilde{v}_i(\sigma, I|j) \right] = v_i(\sigma, I) \tag{6}$$

Proof: $E_{j \sim q_j} \left[\tilde{v}_i(\sigma, I|j) \right]$

$$= \sum_j q_j \tilde{v}_i(\sigma, I|j) \tag{7}$$

$$= \sum_j \sum_{z \in Q_j \cap Z_I} u_i(z) \pi_{-i}^{\sigma}(z[I]), \pi^{\sigma}(z[I], z) \tag{8}$$

$$= \sum_{z \in Z_I} u_i(z) \pi_{-i}^{\sigma}(z[I]), \pi^{\sigma}(z[I], z) \tag{9}$$

$$= \sum_{z \in Z} \sum_{h \in I} u_i(z) \pi_{-i}^{\sigma}(h) \pi^{\sigma}(h, z) \tag{10}$$

$$= v_i(\sigma, I) \tag{11}$$

Equation 9 follows from the fact that $\mathcal{Q}$ is a partition. Equation 10 follows from the fact that $\pi^\sigma(h,z)$ is only non-zero when $h = z[I]$, so only the desired term in the sum will be non-zero. Equation 11 follows from the definition of counterfactual value. ■

Theorem 5 *For any $p \in (0,1]$, if $\forall j \in \{1,\ldots,r\}\, q_j \geq \delta > 0$ at every timestep, then,*

$$R_i^T \leq \left(\frac{1+2\delta+2\sqrt{p}}{\delta\sqrt{p}}\right)\Delta_{u,i}\,|\mathcal{I}_i|\,\sqrt{|A_i|}/\sqrt{T}$$

holds with probability $(1-p)$. Hence, the average strategy profile of two sample-based CFR algorithms in self-play converges to a Nash equilibrium.

Proof: *(Sketch)* [1] We can use Chebyshev's inequality to provide a probabilistic bound on the absolute difference between the sampled counterfactual regret and the true counterfactual regret on a per-information set basis. The bound on this difference contains one term bounding the mean and one term bounding the standard deviation (which depends on δ). Then we show a bound on the expected value of the squared difference between the true and sampled overall counterfactual regret. Using this bound and the triangle inequality we can then bound the sum of the positive counterfactual regret with high probability. Bounding the sum of the positive counterfactual regret in turn bounds the overall regret. ■

3.1 Monte Carlo CFR

We now examine the opposite extreme of vanilla CFR, which we call *Monte Carlo CFR*. In Monte Carlo CFR we choose $\mathcal{Q}$ so that each block contains a single terminal history, *i.e.,* $\forall j \in \{1,\ldots,|Z|\}, |Q_j| = 1$. On each iteration we sample one terminal history, z, 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 $\sigma'_i(I,a) > \epsilon$, then there exists a $\delta > 0$ such that $q_z > \delta$, thus satisfying the conditions of Theorem 5.

This choice of partition results in a simple algorithm. On each iteration, a complete history is sampled using σ' to select actions at each information set, and the product of probabilities of sampling this terminal history is saved. The single history is then traversed forward (to compute each player's probability of playing to reach each prefix of the history, $\pi_i^\sigma(h)$) and backward (to compute each player's probability of playing the remaining actions of the history, $\pi_i^\sigma(h,z)$). During the backward traversal, the sampled counterfactual value for each information set on the path is computed, and the accumulated regret is updated. Finally, the average strategy is updated, and each player's strategy for the next time step σ_i^t is computed from the regrets using Blackwell's algorithm. Since only the regret and policy on the sampled path change, these updates also can be completed on the traversal of the sampled history.

[1] The full proof will be available either as a technical report or part of a future publication.

The choice of the sampling profile is left unspecified in MCCFR. The regret bound suggests the sampling profile should be chosen to make δ as large as possible, implying that uniform random action selection may be the best choice. Intuitively, guiding the sampling towards the outcomes most likely to occur given the current profile also has merit. For this paper we balance these two ideas using an ϵ-greedy exploration, where at each information set we follow the current strategy profile σ^t with probability $(1-\epsilon)$ and choose a uniform random action with probability ϵ. We explore the selection of ϵ in our empirical study.

There are two advantages to MCCFR. First, the cost per iteration is far smaller than vanilla CFR. Vanilla CFR requires $O(|Z|)$ time per iteration while MCCFR requires $O(\ell)$ where ℓ is the length of the longest terminal history. While MCCFR requires more iterations, in the next section we will show empirically that MCCFR's lower cost per iteration often makes up for the required increase in iterations, resulting in faster convergence.

A second advantage is that MCCFR admits a formulation for online regret minimization, where the opponent's strategy is not controlled nor known. If the terminal histories are sampled such that $\sigma'_{-i} = \sigma^t_{-i}$ then we can drop all references to the opponent's strategy σ_{-i} because all the terms cancel with the same terms in σ'_{-i}. In order to minimize regret we would need to choose our own actions so that $\sigma'_i \approx \sigma^t_i$, but with some exploration to guarantee $q_j \geq \delta > 0$. One approach to exploration is to sample a random action with some fixed probability, where this probability is chosen to balance the regret caused by these random actions with the regret from δ being small in the bound. We can then maintain a bound on the average overall regret as long as the number of playings T is known in advance.

4 Experimental Results

Unlike the original CFR work which only was evaluated on abstract poker games, we will examine a collection of games with very different properties. We evaluate the performance of MCCFR on three different games: *One-Card Poker* [Gor05], a slightly modified version of *Goofspiel* [Ros71] where the point card stack is fixed, and *Latent Tic-Tac-Toe*, a version of the classic game where moves are only revealed after the opponent chooses their move. Game sizes are $(|H|,|\mathcal{I}|) = (9N(N-1), 4N)$ for One-Card Poker where N is the deck size, $(98, 3.3)\cdot 10^6$ for modified Goofspiel, and $(70, 8)\cdot 10^6$ for Latent Tic-Tac-Toe.

While all of these games have imperfect information, they represent different types. For example, in One-Card Poker a player's uncertainty consists entirely in the unknown chance outcome. On the other hand, Goofspiel involves no chance. The player's uncertainty is in the opponent's current choice of bid as well as their past bids, of which we only have limited information. In both of these games the ratio of $|H|$ to $|I|$ is large (particularly as the game gets larger), meaning that the players have a high degree of uncertainty. Latent Tic-Tac-Toe, like Goofspiel, has no chance nodes, but the players almost have full information about the state, lacking only their opponent's previous move, a small but critical piece of information. These three games offer diverse settings for equilibria computation.

Our experiments consist of running MCCFR and CFR on the same game and measuring their approximation quality as a function of wall-clock time. Since the two algorithms take radically different amounts of time per iteration, this comparison directly answers if MCCFR's 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), both 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.

Through experimentation we found that contrary to the theoretical bound, an exploration rate ϵ near 1 does not provide the best rate of convergence. An intermediate value of epsilon performs best for Goofspiel. Similar results with similar best choices for epsilon were found in the other domains. The effects of high and low values for epsilon on Goofspiel are displayed in Figure 1; we cannot give a comprehensive analysis of the choice of ϵ here, but the trend is similar in other games. We have also noticed similar trends on the advantage of sampling in MCCFR as $|H|/|\mathcal{I}|$ grows shown for One-Card Poker. Most importantly, MCCFR for a reasonable choice of exploration is finding better approximate equilibria faster than vanilla CFR.

5 Conclusion

In this paper we defined a family of sample-based CFR algorithms for computing approximate equilibria in extensive games, which subsumes all previous CFR variants. We showed that with a reasonable sampling policy, we can bound the average overall regret of any member of this family. We also introduced a new member of this family, Monte Carlo CFR, which samples only a single history for each iteration. We showed that the vast reduction in cost per-iteration can outweigh the increase in the required number of iterations and lead to faster convergence.

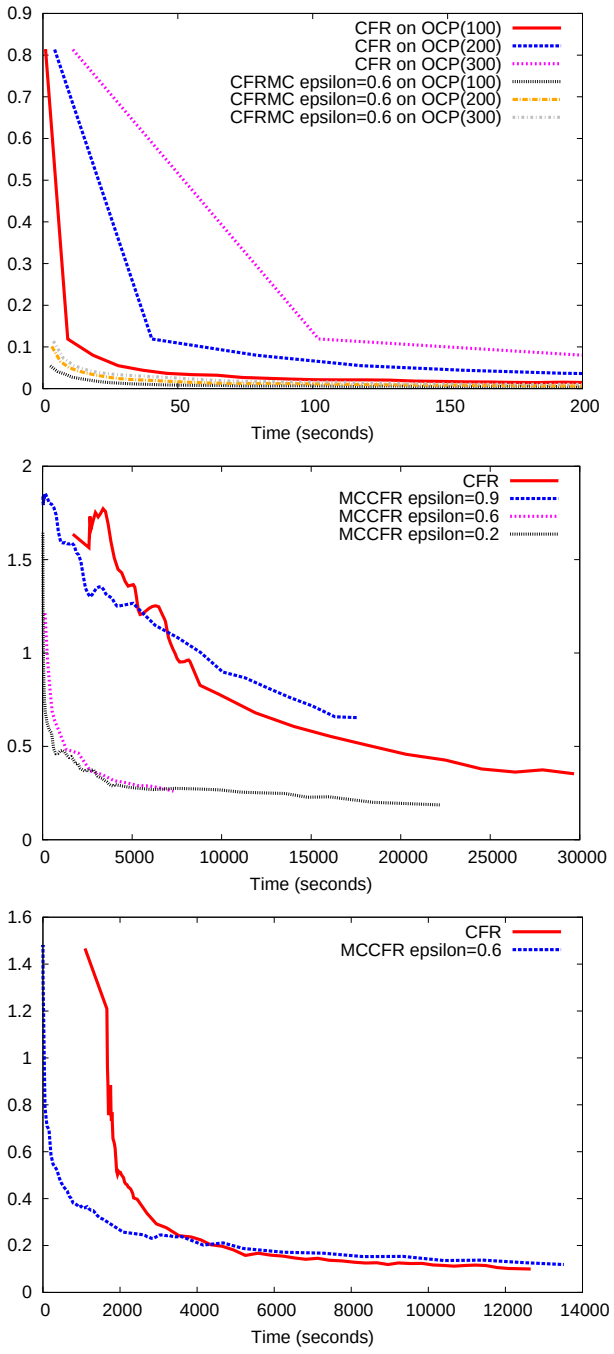

Figure 1: Results for One-Card Poker (top), modified Goofspiel with 7 cards (middle), and Latent Tic-Tac-Toe (bottom). The y-axis in all graphs represents exploitability of the currently approximated ϵ-Nash profile.

References

[Bla56] D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

[GHPS07] 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.

[Gor05] Geoffrey J. Gordon. No-regret algorithms for structured prediction problems. Technical Report CMU-CALD-05-112, Carnegie Mellon University, 2005.

[Gor07] Geoffrey J. Gordon. No-regret algorithms for online convex programs. In *In Neural Information Processing Systems 19*, 2007.

[Ros71] S. M. Ross. Goofspiel — the game of pure strategy. *Journal of Applied Probability*, 8(3):621–625, 1971.

[ZBJP07] Martin Zinkevich, Michael Bowling, Michael Johanson, and Carmelo Piccione. Regret minimization in game with incomplete information. Technical Report TR07-14, University of Alberta, 2007.

[ZJBP08] 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.