Variance Reduction in Monte Carlo Counterfactual Regret Minimization (VR-MCCFR) for Extensive Form Games using Baselines

Martin Schmid1, Neil Burch1, Marc Lanctot1, Matej Moravčík1, Rudolf Kadlec1, Michael Bowling1,2
DeepMind1
University of Alberta2
{mschmid,burchn,lanctot,moravcik,rudolfkadlec,bowlingm}@google.com

Abstract
Learning strategies for imperfect information games from samples of interaction is a challenging problem. A common method for this setting, Monte Carlo Counterfactual Regret Minimization (MCCFR), can have slow long-term convergence rates due to high variance. In this paper, we introduce a variance reduction technique (VR-MCCFR) that applies to any sampling variant of MCCFR. Using this technique, per-iteration estimated values and updates are reformulated as a function of sampled values and state-action baselines, similar to their use in policy gradient reinforcement learning. The new formulation allows estimates to be bootstrapped from other estimates within the same episode, propagating the benefits of baselines along the sampled trajectory; the estimates remain unbiased even when bootstrapping from other estimates. Finally, we show that given a perfect baseline, the variance of the value estimates can be reduced to zero. Experimental evaluation shows that VR-MCCFR brings an order of magnitude speedup, while the empirical variance decreases by three orders of magnitude. The decreased variance allows for the first time CFR+ to be used with sampling, increasing the speedup to two orders of magnitude.

Introduction
Policy gradient algorithms have shown remarkable success in single-agent reinforcement learning (RL) (Mnih et al. 2016; Schulman et al. 2017). While there has been evidence of empirical success in multiagent problems (Foerster et al. 2017; Bansal et al. 2018), the assumptions made by RL methods generally do not hold in multagent partially-observable environments. Hence, they are not guaranteed to find an optimal policy, even with tabular representations in two-player zero-sum (competitive) games (Littman 1995). As a result, policy iteration algorithms based on computational game theory and regret minimization have been the preferred formalism in this setting. Counterfactual regret minimization (Zinkevich et al. 2008) has been a core component of this progress in Poker AI, leading to solving Heads-Up Limit Texas Hold’em (Bowling et al. 2013) and defeating professional poker players in No-Limit (Moravčík et al. 2017; Brown and Sandholm 2017).

This paper similarly unites concepts from both fields, proposing an unbiased variance reduction technique for Monte Carlo counterfactual regret minimization using an analog of state-action baselines from actor-critic RL methods. While policy gradient methods typically involve Monte Carlo estimates, the analog in imperfect information settings is Monte Carlo Counterfactual Regret Minimization (MC-CFR) (Lanctot et al. 2009). Policy gradient estimates based
on a single sample of an episode suffer significantly from variance. A common technique to decrease the variance is a state or state-action dependent baseline value that is subtracted from the observed return. These methods can drastically improve the convergence speed. However, no such methods are known for MCCFR.

MCCFR is a sample based algorithm in imperfect information settings, which approximates counterfactual regret minimization (CFR) by estimating regret quantities necessary for updating the policy. While MCCFR can offer faster short-term convergence than original CFR in large games, it suffers from high variance which leads to slower long-term convergence.

CFR+ provides significantly faster empirical performance and made solving Heads-Up Limit Texas Hold’em possible (Bowling et al. 2015). Unfortunately, CFR+ has so far did not outperform CFR in Monte Carlo settings (Burch 2017) (also see Figure (7) in the appendix for an experiment).

In this work, we reformulate the value estimates using a control variate and a state-action baseline. The new formulation includes any approximation of the counterfactual values, which allows for a range of different ways to insert domain-specific knowledge (if available) but also to design values that are learned online.

Our experiments show two orders of magnitude improvement over MCCFR. For the common tested imperfect information game – Leduc Poker – VR-MCCFR with a state-dependent baseline improved estimates are similar to the ones achieved by MCCFR. Our baseline-improved estimates are similar to the ones determined quality of the states visited (Pepels et al. 2014) and to augment the outcome of a rollout from its length or some predetermined quality of the states visited (Veness et al. 2011).

Our baseline-improved estimates are similar to the ones used in AIVAT (Burch et al. 2018). AIVAT defines estimates of expected values using heuristic values of states as baselines in practice. Unlike this work, AIVAT was only used for evaluation of strategies.

Background

We start with the formal background necessary to understand our method. For details, see (Shoham and Leyton-Brown 2009) (Sutton and Barto 2017).

A two player extensive-form game is tuple $(N, A, H, Z, \tau, u, I)$.

$N = \{1, 2, c\}$ is a finite set of players, where $c$ is a special player called chance. $A$ is a finite set of actions. Players take turns choosing actions, which are composed into sequences called histories; the set of all valid histories is $H$, and the set of all terminal histories (games) is $Z \subseteq H$. We use the notation $A^h = a$ to mean that $a$ is a prefix sequence or equal to $a$. Given a nonterminal history $h$, the player function $\tau: H \times Z \rightarrow N$ determines who acts at $h$. The utility function $u: (N \setminus \{c\}) \times Z \rightarrow [u_{\min}, u_{\max}] \subset \mathbb{R}$ assigns a payoff to each player for each terminal history $z \in Z$.

The notion of a state in imperfect information games requires groupings of histories: $I_i$ for some player $i \in N$ is a partition of $\{h \in H \mid \tau(h) = i\}$ into parts $I_i \subseteq H$ such that $h, h' \in I_i$ if player $i$ cannot distinguish $h'$ from $h'$ given the information known to player $i$ at two histories. We call these information sets. For example, in Texas Hold’em poker, for all $I \in I_i$, the (public) actions are the same for all $h, h' \in I$, and $h$ only differs from $h'$ in cards dealt to the opponents (actions chosen by chance). For convenience, we refer to $I(h)$ as the information state that contains $h$.

At any $I$, there is a subset of legal actions $A(I) \subseteq A$. To choose actions, each player $i$ uses a strategy $\sigma_i: I \rightarrow \Delta(A(I))$, where $\Delta(X)$ refers to the set of probability distributions over $X$. We use the shorthand $\sigma(h, a)$ to refer to $\sigma(I(h), a)$. Given some history $h$, we define the reach probability $\pi^a(h) = \Pi_{h' \in A^h} \sigma(\tau(h'))(I(h'), a)$ to be the product of all action probabilities leading up to $h$. This reach probability contains all players’ actions, but can be separated into $\pi^a(h) = \sigma^a(h)\pi^a(\tau(h))$ into player $i$’s actions’ contribution and the contribution of the opponents’ of player $i$ (including chance).

Finally, it is often useful to consider the augmented information sets (Burch et al. 2014). While an information set $I$ groups histories $h$ that player $i = \tau(h)$ cannot distinguish,
an augmented information set histories that player \( i \) can not distinguish, including these where \( \tau(h) \neq i \). For a history \( h \), we denote an augmented information set of player \( i \) as \( I_i(h) \). Note that the if \( \tau(h) = i \) then \( I_i(h) = I(h) \) and \( I(h) = I_{\tau(h)}(h) \).

**Counterfactual Regret Minimization**

Counterfactual Regret (CFR) Minimization is an iterative algorithm that produces a sequence of strategies \( \sigma^0, \sigma^1, \ldots, \sigma^T \), whose average strategy \( \bar{\sigma}^T \) converges to an approximate Nash equilibrium as \( T \to \infty \) in two-player zero-sum games (Zinkevich et al. 2008). Specifically, on iteration \( t \), for each \( I \), it computes **counterfactual values**.

Define \( Z_I = \{(h,z) \in H \times Z | h \in I, h \not\in z\} \), and \( u^T_i(h,z) = \pi^\sigma(h,z)u_i(z) \). We will also sometimes use the short form \( u^T_i(h) = \sum_{z \in Z, h \not\in z} u^T_i(h,z) \). A counterfactual value is:

\[
v_i(\sigma^t, I) = \sum_{(h,z) \in Z_I} \pi_{h,z}^\sigma(h)u^T_i(h,z).
\]

(1)

We also define an action-dependent counterfactual value,

\[
v_i(\sigma, I, a) = \sum_{(h,z) \in Z_I} \pi_{h,z}^\sigma(h,a)u(\sigma, h, a),
\]

(2)

where \( ha \) is the sequence \( h \) followed by the action \( a \). The values are analogous to the difference in \( Q \)-values and \( V \)-values in RL, and indeed we have \( v_i(\sigma, I) = \sum_a v_i(\sigma, I, a) \). CFR then computes a **counterfactual regret** for not taking \( a \) at \( I \):

\[
r^T(I, a) = v_i(\sigma^t, I) - v_i(\sigma^t, I, a),
\]

(3)

This regret is then accumulated \( R^T(I, a) = \sum_{t=1}^T r^T(I, a) \), which is used to update the strategies using **regret-matching** (Hart and Mas-Colell 2000):

\[
\sigma^{t+1}(I) = \frac{(R^T(I, a))^+}{\sum_{a \in A(I)} (R^T(I, a))^+},
\]

(4)

where \( \{x\}^+ = \max(x, 0) \), or to the uniform strategy if \( \sum_a (R^T(I, a))^+ = 0 \). CFR+ works by thresholding the quantity at each round (Tammelin et al. 2015): define \( Q^0(I, a) = 0 \) and \( Q^T(I, a) = (Q^{T-1} + r^T(I, a))^+ \); CFR+ updates the policy by replacing \( R^T \) by \( Q^T \) in equation \( 4 \). In addition, it always alternates the regret updates of the players (whereas some variants of CFR update both players), and the average strategy places more (linearly increasing) weight on more recent iterations.

If for player \( i \) we denote \( u_i(\sigma) = u_i(\sigma, \sigma_{-i}) \), and run CFR for \( T \) iterations, then we can define the **overall regret** of the strategies produced as:

\[
R^T_i = \max_{\sigma_i} \sum_{t=1}^T (v_i(\sigma_i^t, \sigma_{-i}^t) - v_i(\sigma^t)).
\]

CFR ensures that \( R^T_i / T \to 0 \) as \( T \to \infty \). When two players minimize regret, the folk theorem then guarantees a bound on the distance to a Nash equilibrium as a function of \( R^T_i / T \).

To compute \( v_i \) precisely, each iteration requires traversing over subtrees under each \( a \in A(I) \) at each \( I \). Next, we describe variants that allow sampling parts of the trees and using estimates of these quantities.

**Monte Carlo CFR**

Monte Carlo CFR (MCCFR) introduces sample estimates of the counterfactual values, by visiting and updating quantities over only part of the entire tree. MCCFR is a general family of algorithms: each instance defined by a specific sampling policy. For ease of exposition and to show the similarity to RL, we focus on outcome sampling (Lanctot et al. 2009), however, our baseline-enhanced estimates can be used in all MCCFR variants. A **sampling policy** \( \xi \) is defined in the same way as a strategy (a distribution over \( A(I) \) for all \( I \)) with a restriction that \( \xi(h, a) > 0 \) for all histories and actions. Given a terminal history sampled with probability \( q(z) = \pi^\xi(z) \), a **sampled counterfactual value** \( \hat{v}_i(\sigma, I, z) \)

\[
\hat{v}_i(\sigma, I, z) = \frac{\pi^\sigma(h)u_i(h, z)}{q(z)}, \quad \text{for } h \in I, h \not\in z,
\]

(5)

and 0 for histories that were not played, \( h \not\in z \). The estimate is unbiased: \( \mathbb{E}_{z \sim \xi} [\hat{v}_i(\sigma, I, z)] = v_i(\sigma, I) \), by (Lanctot et al. 2009; Lemma 1). As a result, \( \hat{v}_i \) can be used in Equation 3 to accumulate estimated regrets \( \hat{r}^T(I, a) = \hat{v}_i(\sigma^t, I, a) - \hat{v}_i(\sigma^t, I) \) instead. The regret bound requires an additional term \( \frac{1}{\min_{z \in Z} q(z)} \), which is exponential in the length of \( z \) and similar observations have been made in RL (Arjona-Medina et al. 2018). The main problem with the sampling variants is that they introduce variance that can have a significant effect on long-term convergence (Gibson et al. 2012).

**Control Variates**

Suppose one is trying to estimate a statistic of a random variable, \( X \), such as its mean, from samples \( X = (X_1, X_2, \ldots, X_n) \). A crude Monte Carlo estimator is defined to be \( \hat{X}^{mc} = \frac{1}{n} \sum_{i=1}^n X_i \). A control variate is a random variable \( Y \) with a known mean \( \mu_Y = \mathbb{E}[Y] \), that is paired with the original variable, such that samples are instead of the form \((X, Y)\) (Owen 2013). A new random variable is then defined, \( Z_i = X_i + c(Y_i - \mu_Y) \). An estimator \( \hat{Z}^{cv} = \frac{1}{n} \sum_{i=1}^n Z_i \). Since \( \mathbb{E}[Z_i] = \mathbb{E}[X_i] \) for any value of \( c \), \( \hat{Z}^{cv} \) can be used in place of \( \hat{X}^{mc} \) with variance \( \mathbb{V}[Z_i] = \mathbb{V}[X_i] + c^2 \mathbb{V}[Y_i] + 2c \mathbb{C}[X_i, Y_i] \). So when \( X \) and \( Y \) are positively correlated and \( c < 0 \), variance is reduced when \( \mathbb{C}[X, Y] > \frac{c^2}{2} \mathbb{V}[Y] \).

**Reinforcement Learning Mapping**

There are several analogies to make between Monte Carlo CFR in imperfect information games and reinforcement learning. Since our technique builds on ideas that have been widely used in RL, we end the background by providing a small discussion of the links.

First, dynamics of an imperfect information game are similar to a partially-observable episodic MDP without any cycles. Policies and strategies are identically defined, but in imperfect information games a deterministic optimal (Nash) strategy may not exist causing most of the RL methods to fail to converge. The search for a minmax-optimal strategy with several players is the main reason CFR is used instead of, for example, value iteration. However, both operate by defining values of states which are analogous (counterfactual values
versus expected values) since they are both functions of the strategy/policy; therefore, can be viewed as a kind of policy iteration which computes the values and from which a policy is derived. However, the iterates $\sigma^i$ are not guaranteed to converge, only the average strategy $\hat{\sigma}$ does.

Monte Carlo CFR is an off-policy Monte Carlo analog. The value estimates are unbiased specifically because they are corrected by importance sampling. Most applications of MCCFR have operated with tabular representations, but this is mostly due to the differences in objectives. Function approximation methods have been proposed for CFR (Waugh et al. 2015) but the variance from pure Monte Carlo methods may prevent such techniques in MCCFR. The use of baselines has been widely successful in policy gradient methods, so reducing the variance could enable the practical use of function approximation in MCCFR.

It was recently shown that policy gradient and actor-critic algorithms implement a form of function approximation in MCCFR.

The proof is in the appendix. Since each estimate builds on other estimates, the benefit of the reduction in variance can be propagated up through the tree.

Another key result is that there exists a perfect baseline that leads to zero-variance estimates at the updated information sets.

Monte Carlo CFR with Baselines

We now introduce our technique: MCCFR with baselines. While the baselines are analogous to those from policy gradient methods (using counterfactual values), there are slight differences in their construction.

Our technique constructs value estimates using control variates. Note that MCCFR is using sampled estimates of counterfactual values $\check{v}_i(\sigma, I)$ whose expected value is the counterfactual value $v_i(\sigma, I)$. First, we introduce an estimated counterfactual value $\hat{v}_i(\sigma, I)$ to be any estimator of the counterfactual value (not necessarily $\check{v}_i$ as defined above, but this is one possibility).

We now define an action-dependent baseline $b_i(I, a)$ that, as in RL, serves as a basis for the sampled values. The intent is to define a baseline function to approximate or be correlated with $E[\check{v}_i(\sigma, I, a)]$. We also define a sampled baseline $\hat{b}_i(I, a)$ as an estimator such that $E[\hat{b}_i(I, a)] = b_i(I, a)$. From this, we construct a new baseline-enhanced estimate for the counterfactual values:

$$\hat{v}_i^b(\sigma, I, a) = \check{v}_i(\sigma, I, a) - \hat{b}_i(\sigma, I, a) + b_i(\sigma, I, a)$$ (6)

First, note that $\hat{b}_i$ is a control variate with $c = -1$. Therefore, it is important that $\hat{b}_i$ be correlated with $\check{v}_i$. The main idea of our technique is to replace $\check{v}_i(\sigma, I, a)$ with $\hat{v}_i^b(\sigma, I, a)$. A key property is that by doing so, the expectations remain unchanged.

Lemma 1. For any $i \in \mathcal{N} - \{c\}, \sigma_i, I \in \mathcal{I}, a \in A(I)$, if $E[\hat{b}_i(I, a)] = b_i(I, a)$ and $E[\check{v}_i(\sigma, I, a)] = v_i(\sigma, I, a)$, then $E[\hat{v}_i^b(\sigma, I, a)] = v_i(\sigma, I, a)$.

The proof is in the appendix. As a result, any baseline whose expectation is known can be used and the baseline-enhanced estimates are consistent. However, not all baselines will decrease variance. For example, if $\text{Cov}[\check{v}_i, \hat{b}_i]$ is too low, then the $\text{Var}[\hat{b}_i]$ term could dominate and actually increase the variance.

Recursive Bootstrapping

Consider the individual computation (1) for all the information sets on the path to a sampled terminal history $z$. Given that the counterfactual values up the tree can be computed from the counterfactual values down the tree, it is natural to consider propagating the already baseline-enhanced counterfactual values (6) rather than the original noisy sampled values - thus propagating the benefits up the tree. The Lemma (2) then shows that by doing so, the updates remain unbiased. Our experimental section shows that such bootstrapping a crucial component for the proper performance of the method.

To properly formalize this bootstrapping computation, we must first recursively define the expected value:

$$\hat{u}_i(\sigma, h, a|z) = \begin{cases} \hat{u}_i(\sigma, ha|z)/\xi(h, a) & \text{if } ha \subseteq z \\ 0 & \text{otherwise} \end{cases}$$ (7)

and

$$\hat{u}_i(\sigma, h|z) = \begin{cases} u_i(h) & \text{if } h = z \\ \sum_a \sigma(h, a)\hat{u}_i(\sigma, h, a|z) & \text{if } h \subseteq z \\ 0 & \text{otherwise} \end{cases}$$ (8)

Next, we define a baseline-enhanced version of the expected value. Note that the baseline $b_i(I, a)$ can be arbitrary, but we discuss a particular choice and update of the baseline in the later section. For every action, given a specific sampled trajectory $z$, then $\hat{u}_i^b(\sigma, h, a|z) = \hat{u}_i(\sigma, h, a|z) - \hat{b}_i(I(h), a)$.

$$\hat{u}_i^b(\sigma, h, a|z) = \begin{cases} b_i(I(h), a) + \hat{u}_i^b(\sigma, ha|z) - b_i(I(h), a) & \text{if } ha \subseteq z \\ 0 & \text{if } h \subseteq z, ha \nsubseteq z \text{ otherwise} \end{cases}$$ (9)

and

$$\hat{u}_i^b(\sigma, h|z) = \begin{cases} u_i(h) & \text{if } h = z \\ \sum_a \sigma(h, a)\hat{u}_i^b(\sigma, h, a|z) & \text{if } h \subseteq z \\ 0 & \text{otherwise} \end{cases}$$ (10)

These are the values that are bootstrapped. We estimate counterfactual values needed for the regret updates using these values as:

$$\hat{v}_i^b(\sigma, I(h), a|z) = \hat{v}_i^b(\sigma, h, a|z) = \frac{\sigma^*_{e_i}(h)}{q(h)}\hat{u}_i^b(\sigma, h, a|z).$$ (11)

We can now formally state that the bootstrapping keeps the counterfactual values unbiased:

Lemma 2. Let $\hat{v}_i^b$ be defined as in Equation (11). Then, for any $i \in \mathcal{N} - \{c\}, \sigma_i, I \in \mathcal{I}, a \in A(I)$, it holds that $E_z[\hat{v}_i^b(\sigma, I, a|z)] = v_i(\sigma, I, a)$.

The proof is in the appendix. Since each estimate builds on other estimates, the benefit of the reduction in variance can be propagated up through the tree.

Another key result is that there exists a perfect baseline that leads to zero-variance estimates at the updated information sets.
How does one choose a baseline, given that we want these to be good estimates of the individual counterfactual values? A common choice of baseline in policy gradient algorithms is the mean value of the state, which is learned online (Mnih et al. 2016). Inspired by this, we choose a similar quantity: the average expected value 

\[ \bar{u}_i(I, a) = \frac{1}{|I|} \sum_{a} \bar{v}_i(I, a) \] 

where \( \bar{v}_i(I, a) \) is the average value of \( v_i(I, a) \) over all visits to \( I \) at iteration \( t \).

We then define the baseline \( b_i(I, a) = \bar{u}_i(I, a) \), and

\[ \hat{b}_i(I, a|z) = \begin{cases} b_i(I, a) / \xi(I, a) & \text{if } h_a \subseteq z, h_e \in I \setminus z \text{ is terminal} \\ 0 & \text{otherwise} \end{cases} \] 

The baseline can therefore be thought of as \textit{local} to \( I \), since it depends only on quantities defined and tracked at \( I \). Note that \( E_{a \sim \xi(I, a)}[\hat{b}_i(I, a|z)] = b_i(I, a) \) as required.

**Choice of Baselines**

How does one choose a baseline, given that we want these to be good estimates of the individual counterfactual values? A common choice of the baseline in policy gradient algorithms is the mean value of the state, which is learned online (Mnih et al. 2016). Inspired by this, we choose a similar quantity: the average expected value \( \bar{u}_i(I, a) \). That is, in addition to accumulating regret for each \( I \), average expected values are also tracked.

While a direct average can be tracked, we found that an exponentially-decaying average that places heavier weight on more recent samples to be more effective in practice. On the \( k^{th} \) visit to \( I \) at iteration \( t \),

\[ \tilde{u}_i(I, a) = \begin{cases} 0 & \text{if } k = 0 \\ (1 - \alpha) \tilde{u}_i^{k-1}(I, a) + \alpha \tilde{u}_i^0(I, a)^* & \text{if } k > 0 \end{cases} \] 

We then define the baseline \( b_i(I, a) = \tilde{u}_i(I, a) \), and

\[ \hat{b}_i(I, a|z) = \begin{cases} b_i(I, a) / \xi(I, a) & \text{if } h_a \subseteq z, h_e \in I \setminus z \text{ is terminal} \\ 0 & \text{otherwise} \end{cases} \] 

The baseline can therefore be thought of as \textit{local} to \( I \), since it depends only on quantities defined and tracked at \( I \). Note that \( E_{a \sim \xi(I, a)}[\hat{b}_i(I, a|z)] = b_i(I, a) \) as required.

**Summary of the Full Algorithm**

We now summarize the technique developed above. One iteration of the algorithm consists of:

1. Repeat the steps below for each \( i \in \mathcal{N} \setminus \{c\} \).
2. Sample a trajectory \( z \sim \xi \).
3. For each history \( h \subseteq z \) in reverse order (longest first):
   (a) If \( h \) is terminal, simply return \( u_i(h) \).
   (b) Obtain current strategy \( \sigma(I) \) from Eq. 9 using cumulative regrets \( R(I, a) \) where \( h \in I \).
   (c) Use the child value \( \hat{v}_i^k(\sigma, ha) \) to compute \( \hat{u}_i^k(\sigma, h) \) as in Eq. 9.
   (d) If \( \pi(h) = i \) then for \( a \in A(I) \), compute \( \hat{v}_i^k(\sigma, I, a) = \frac{\pi(h)}{q(h)} \hat{u}_i^k(\sigma, ha) \) and accumulate regrets \( R(I, a) \leftarrow R(I, a) + \hat{v}_i^k(\sigma, I, a) - \hat{v}_i^k(\sigma, I) \).
   (e) Update \( \hat{u}_i(\sigma, I, a) \).
   (f) Finally, return \( \hat{u}_i^k(\sigma, h) \).

Note that the original outcome sampling is an instance of this algorithm. Specifically, when \( \hat{b}_i(I, a) = 0 \), then \( \hat{v}_i^k(\sigma, I, a) = \hat{v}_i(\sigma, I, a) \). Step by step example of the computation is in the appendix.

**Experimental Results**

We evaluate the performance of our method on Leduc poker (Southey et al. 2005), a commonly used benchmark poker game. Players have an unlimited number of chips, and the deck has six cards, divided into two suits of three identically-ranked cards. There are two rounds of betting: after the first round a single public card is revealed from the deck. Each player antes 1 chip to play, receiving one private card. There are at most two bet or raise actions per round, with a fixed size of 2 chips in the first round, and 4 chips in the second round.

For the experiments, we use a vectorized form of CFR that applies regret updates to each information set consistent with the public information. The first vector variants were introduced in (Johanson et al. 2012), and have been used in DeepStack and Libratus (Moravčík et al. 2017) (Brown and Sandholm 2017). See the appendix for more detail on the implementation. Baseline average values \( \hat{u}_i^k(I, a) \) used a
decay factor of $\alpha = 0.5$. We used a uniform sampling in all our experiments, $\xi(I, a) = \frac{1}{|A(I)|}$.

We also consider the best case performance of our algorithm by using the oracle baseline. It uses baseline values of the true counterfactual values. We also experiment with and without CFR+, demonstrating that our technique allows the CFR+ to be for the first time efficiently used with sampling.

**Convergence**

We compared MCCFR, MCCFR+, VR-MCCFR, VR-MCCFR+, and VR-MCCFR+ with the oracle baseline, see Fig. 3. The variance-reduced VR-MCCFR and VR-MCCFR+ variants converge significantly faster than plain MCCFR. Moreover, the speedup grows as the baseline improves during the computation. A similar trend is shown by both VR-MCCFR and VR-MCCFR+, see Fig. 4. MCCFR needs hundreds of millions of iterations to reach the same exploitability as VR-MCCFR+ achieves in one million iterations: a 250-times speedup. VR-MCCFR+ with the oracle baseline significantly outperforms VR-MCCFR+ at the start of the computation, but as time progresses and the learned baseline improves, the difference shrinks. After one million iterations, exploitability of VR-MCCFR+ with a learned baseline approaches the exploitability of VR-MCCFR+ with the oracle baseline. This oracle baseline result gives a bound on the gains we can get by constructing better learned baselines.

**Observed Variance**

To verify that the observed speedup of the technique is due to variance reduction, we experimentally observed variance of counterfactual value estimates for MCCFR+ and MCCFR, see Fig. 5. We did that by sampling 1000 alternative trajectories for all visited information sets, with each trajectory sampling a different estimate of the counterfactual value. While the variance of value estimates in the plain algorithm seems to be more or less constant, the variance of VR-MCCFR and VR-MCCFR+ value estimates is lower, and continues to decrease as more iterations are run. This confirms that the combination of baseline and bootstrapping is reducing variance, which implies better performance given the connection between variance and MCCFR’s performance (Theorem 1).

**Evaluation of Bootstrapping and Baseline Dependence on Actions**

Recent work that evaluates action-dependent baselines in RL (Tucker et al. 2018), shows that there is often no real advantage compared to baselines that depend just on the state. It is also not common to bootstrap the value estimates in RL. Since VR-MCCFR uses both of these techniques it is natural to explore the contribution of each idea. We compared four VR-MCCFR+ variants: with or without bootstrapping and with baseline that is state or state-action dependant, see Fig. 6. The conclusion is that the improvement in the performance is very small unless we use both bootstrapping and an action-dependant baseline.

![Figure 3: Convergence of exploitability for different MCCFR variants on logarithmic scale. VR-MCCFR converges substantially faster than plain MCCFR. VR-MCCFR+ bring roughly two orders of magnitude speedup. VR-MCCFR+ with oracle baseline (actual true values are used as baselines) is used as a bound for VR-MCCFR’s performance to show possible room for improvement. When run for $10^6$ iterations VR-MCCFR+ approaches performance of the oracle version. The ribbons show 5th and 95th percentile over 100 runs.](image1)

![Figure 4: Speedup of VR-MCCFR and VR-MCCFR+ compared to plain MCCFR. Y-axis show how many times more iterations are required by MCCFR to reach the same exploitability as VR-MCCFR or VR-MCCFR+.](image2)

**Conclusions**

We have presented a new technique for variance reduction for Monte Carlo counterfactual regret minimization. This technique has close connections to existing RL methods of
Using this technique, we show that empirical variance is in-


gestion that state-action baselines are superior to state baselines. In contrast to RL environments, our experiments in imperfect information games suggest that state-action baselines are superior to state baselines. Using this technique, we show that empirical variance is in-

deed reduced, speeding up the convergence by an order of magnitude. The decreased variance allows for the first time CFR+ to be used with sampling, bringing the speedup to two orders of magnitude. Finally, the technique requires only a relatively small computational overhead. In the experiments on Leduc using our non-optimized implementation, we observed a factor of 2 per-iteration slowdown.

References


Appendices

MCCFR and MCCFR+ comparison

While it is known that MCCFR+ is outperformed by MCCFR (Burch 2017), we are not aware of any explicit comparison of these two algorithms in literature. Fig. 7 shows experimental evaluation of these two techniques on Leduc poker.

![Convergence on Leduc poker](image)

Figure 7: Convergence of MCCFR and MCCFR+ on logarithmic scale. For the first $10^0$ iterations, MCCFR+ performs similarly to the MCCFR. After approximately $10^7$ iterations, the difference in favor of MCCFR starts to be visible and the gap in exploitability widens as the number of iterations grows.

Vector Form CFR and Augmented Information Sets

The first appearance of the vector form was presented in (Johanson et al. 2011). In this paper, the best response computation, needed to compute exploitability, was sped-up by redefining the computation using the notion of a public tree. At the heart of a public tree is the notion of a public state which contains a set of information sets whose histories are consistent with the public information revealed so far (Johanson et al. 2011) Definition 2). This allowed the method to compute quantities for all information sets consistent with a public state at once (stored in vectors) and operations to compute them could be vectorized during a traversal of the public tree. There are also game-specific optimizations that could be applied at leaf nodes to asymptotically reduce the total computation necessary.

A similar construction was used in several sampling variants introduced in (Johnson et al. 2012). Here, instead of computing necessary for best response, counterfactual values were vectorized and stored instead. The paper describes several ways to sample at various types of chance nodes (ones which reveal public information, or private information to each player), but the concept of a vectorized form of CFR was general. In fact, a vector form of vanilla CFR is possible in any game: when traversing down the tree, these vectors store the probability of reaching each information set (called a range in (Moravčík et al. 2017)) and return vectors of counterfactual values. Both DeepStack and Libratus used vector forms of CFR and CFR+ in No-Limit poker.

Our experiments use a MCCFR variant of the public tree CFR methods above. Given some sampled sequence of public actions, we update all information sets consistent with that sequence. For example in Leduc poker, six trajectories per player are considered, one for every possible private player card, which all share the same sequence of public actions.

In this public tree form, it is often useful to consider the counterfactual values for the player that is not acting. For example, in poker it might be my turn to act, but I want a baseline value estimate of how well my opponent would do with any possible hand, given the publicly visible betting and board cards. Whereas standard information sets only partition states according to the acting player, augmented information sets (Burch et al. 2014) partition the states according to the player and that player’s information. Two states are said to be in the same player $p$ augmented information set if both states passed through the same sequence of player $p$ information sets and made the same player $p$ actions. In poker, these are states where player $p$’s private card and the public tree node (the betting and board cards) are the same.

In our VR-MCCFR implementation, the baselines values are kept as vectors at each public state, each representing a baseline for the augmented information sets corresponding to the public state. Also, the average values tracked are counterfactual and normalized by the range. So, for example in Leduc, for five information sets in some public state, $(I_1, I_2, ..., I_5)$, quantity tracked by the baseline at this public state for action $a$ is:

$$\hat{v}_i(\sigma, I_k, a) \sum_{k'} \pi^p_{opp}(I_{k'}^{opp})$$

where $\pi^p_{opp}$ is the reach probability of the opponent only (excluding chance), and $I^{opp}_k$ refers to the augmented information set belonging to the opponent at $I$. Then, when using the baseline values to compute the modified counterfactual values, we need to multiply them by the current $\sum_{k'} \pi^p_{opp}(I_{k'}^{opp})$ to get the baseline values under the current strategy $\sigma$.

Proofs

Proof of Lemma 1

$$\mathbb{E}[\hat{v}(\sigma, I, a)] = \mathbb{E}[\hat{v}_i(\sigma, I, a)] = \mathbb{E}[\hat{b}_i(I, a)] + \mathbb{E}[b_i(I, a)]$$

$$= v_i(\sigma, I, a) - b_i(I, a) + b_i(I, a)$$

$$= v_i(\sigma, I, a).$$

Proof of Lemma 2

We begin by proving a few supporting lemmas regarding local expectations over actions at specific histories:

Lemma 4. Given some $h \in \mathcal{H}$, for any $z \in \mathcal{Z}$ generated by sampling $\xi : \mathcal{H} \rightarrow \mathcal{A}$ and all actions $a$, $\mathbb{E}_{z \sim \xi}[\hat{u}_i(\sigma, h, a, z)] = \sum_{z, h a \in z} q(z) \hat{u}_i(\sigma, h a | z) / \xi(h, a)$;
Lemma 5. Given some $h \in \mathcal{H}$, for any $z \in \mathcal{Z}$ generated by sampling $\xi : \mathcal{H} \mapsto \mathcal{A}$, the local baseline-enhanced estimate is an unbiased estimate of expected values for all actions $a$:
\[
E_{z \sim \xi}[\hat{u}_i^b(\sigma, h, a|z)] = E_{z \sim \xi}[\hat{u}_i(\sigma, h, a|z)].
\]

Proof. We prove this by induction on the maximum distance from $ha$ to any terminal. The base case is $ha \in \mathcal{Z}$. 
\[
E_{z \sim \xi}[\hat{u}_i^b(\sigma, h, a|z)] = \sum_{z, ha \in \mathcal{Z}} q(z) \hat{u}_i^b(\sigma, ha|z)/\xi(h, a) \quad \text{by Lemma 4}
\]
\[
= \sum_{z, ha \in \mathcal{Z}} q(z) \hat{u}_i(\sigma, ha|z)/\xi(h, a) \quad \text{by Eq. 10}
\]
\[
= E_{z \sim \xi}[\hat{u}_i(\sigma, h, a|z)] \quad \text{by Eq. 7 and Lemma 6}.
\]

Now assume for $i \geq 0$ that the lemma property holds for all $h'$ at most $j \leq i$ steps from a terminal. Consider history $ha$ being $i+1$ steps from some terminal, which implies that $ha \not\in \mathcal{Z}$. We have 
\[
E_{z \sim \xi}[\hat{u}_i^b(\sigma, h, a|z)] = \sum_{z, ha \in \mathcal{Z}} q(z) \hat{u}_i^b(\sigma, ha|z)/\xi(h, a) \quad \text{by Lemma 4}
\]
\[
= \sum_{z, ha \in \mathcal{Z}} q(z) \sum_{a'} \sigma(ha, a') \hat{u}_i^b(\sigma, ha, a'|z)/\xi(h, a) \quad \text{by assumption}
\]
\[
= \sum_{z, ha \in \mathcal{Z}} q(z) \hat{u}_i(\sigma, ha|z)/\xi(h, a) \quad \text{by Eq. 8}
\]
\[
= E_{z \sim \xi}[\hat{u}_i(\sigma, h, a|z)] \quad \text{by Eq. 7}.
\]

The lemma property holds for distance $i+1$, and so by induction the property holds for all $h$ and $a$.


Lemma 6. Given some $h \in \mathcal{H}$, for any $z \in \mathcal{Z}$ generated by sampling $\xi : \mathcal{H} \mapsto \mathcal{A}$ and for all actions $a$, the local baseline-enhanced estimate is an unbiased estimate of the original sampled counterfactual value:
\[
E_{z \sim \xi}[\hat{v}_i^b(\sigma, I_i(h), a|z)] = E_{z \sim \xi}[\hat{v}_i(\sigma, I_i(h), a|z)].
\]

Proof. First, 
\[
E_{z \sim \xi}[\hat{v}_i^b(\sigma, I_i(h), a|z)] = E_{z \sim \xi}[\hat{v}_i(\sigma, I_i(h), a|z)]
\]
\[
= \frac{\pi_i^\sigma(h)}{\pi_i(h)} E_{z \sim \xi}[\hat{v}_i^b(\sigma, h, a|z)] \quad \text{by Eq. 11}
\]
\[
= \frac{\pi_i^\sigma(h)}{\pi_i(h)} \frac{\pi_i^\sigma(h)}{\pi_i(h)} E_{z \sim \xi}[\hat{v}_i(\sigma, h, a|z)] \quad \text{by Lemma 5}
\]
\[
= E_{z \sim \xi}[\hat{v}_i(\sigma, I_i(h), a|z)] \quad \text{by Eq. 5 and Lemma 6}.
\]

Proof of Lemma 2. The proof now follows directly:
\[
E_{z \sim \xi}[\hat{v}_i^b(\sigma, I_i(h), a|z)] = E_{z \sim \xi}[\hat{v}_i(\sigma, I_i(h), a|z)] = v_i(\sigma, I_i, a) \quad \text{by (Lanctot et al. 2009) Lemma 1}.
\]

Proof of Lemma 3.

We start by proving that given an oracle baseline, the baseline-enhanced expected value is always equal to the true expected value, and therefore has zero variance.

Lemma 7. Using an oracle baseline defined over histories, $b_i^\sigma(h, a) = u_i^\sigma(ha)$, then for all $z$ such that $h \subseteq z$, $\hat{u}_i^b(\sigma, h, a|z) = u_i^\sigma(ha)$.

Proof. Similar to above, we prove this by induction on the maximum distance from $ha$ to $z$. The base case is $ha \in \mathcal{Z}$.
\[
E_{z \sim \xi}[\hat{u}_i^b(\sigma, h, a|z)] = \sum_{z, ha \in \mathcal{Z}} q(z) \hat{u}_i^b(\sigma, ha|z)/\xi(h, a) \quad \text{by Lemma 4}
\]
\[
= \sum_{z, ha \in \mathcal{Z}} q(z) \hat{u}_i(\sigma, ha|z)/\xi(h, a) \quad \text{by Eq. 10 and definition of $b_i^\sigma(h, a)$}
\]
\[
= \sum_{z, ha \in \mathcal{Z}} q(z) \hat{u}_i(\sigma, ha|z)/\xi(h, a) \quad \text{by Eq. 8}
\]
\[
= E_{z \sim \xi}[\hat{u}_i(\sigma, h, a|z)] \quad \text{by Eq. 7}.
\]

The lemma property holds for distance $i + 1$, and so by induction the property holds for all $h$ and $a$. 

\[
\hat{u}_i^b(\sigma, ha|z) = u_i^\sigma(ha) \quad \text{(12)}
\]

because $\hat{u}_i^b(\sigma, ha|z)$
\[
= \sum_{a'} \sigma(ha, a') \hat{u}_i^b(\sigma, ha, a'|z) \quad \text{by Eq. 10}
\]
\[
= \sum_{a'} \sigma(ha, a') u_i^\sigma(haa') \quad \text{by assumption}
\]
\[
= u_i^\sigma(ha) \quad \text{by definition of $u_i^\sigma$}.
\]
We now look at $\check{u}^*_i (\sigma, h, a | z) \equiv \{ u^\sigma_i (ha) + \frac{u^\sigma_i (\sigma, ha | z) - \xi(h, a)}{\xi(h, a)} u^\sigma_i (ha) }$ if $ha \sqsubseteq z \quad \text{otherwise}$

by Eq. 9 and definition of $b^*_i (h, a)$

$= \{ u^\sigma_i (ha) + \frac{u^\sigma_i (ha) - \xi(h, a)}{\xi(h, a)} u^\sigma_i (ha) \}$ if $ha \sqsubseteq z \quad \text{otherwise}$

by Eq. 11

$= u^\sigma_i (ha)$

The lemma property holds for distance $i + 1$, and so by induction the property holds for all $h$ and $a$.

Proof of Lemma: Given $z$ such that $h \sqsubseteq z$, we have $\check{v}^*_i (\sigma, h, a | z) = \frac{\pi^\sigma_i (h)}{q(h)} u^\sigma_i (ha)$ by Lemma 7.

None of the terms above depend on $z$, and so we have $\forall h, z \sim \xi, h \subseteq z [\check{v}^*_i (\sigma, h, a | z)] = 0$. Note as well that $\pi^\sigma_i (h) u^\sigma_i (ha)$ corresponds to the terms in the summation of Equation 2; so abusing notation, we have $\check{v}^*_i (\sigma, h, a | z) = v_i (\sigma, h, a) / q(h)$. The counterfactual value of taking action $a$ at $h$, with an importance sampling weight to correct for the likelihood of reaching $h$.

In MCCFR, the optimal baseline $b^*$ is not known, as it would require traversing the entire tree, taking away any advantages of sampling. However, $b^*$ can be approximated (learned online), which motivates the choice for tracking its average value presented in the main part of the paper.

Kuhn Example

In this section, we present a step-by-step example of one iteration of the algorithm on Kuhn poker (Kuhn poker 2018). Kuhn poker is a simplified version of poker with three cards and is therefore suitable for demonstration purposes. Table 1 shows forward pass of VR-MCCFR algorithm, Table 2 shows backward pass.
Table 1: Detailed example of updates computed for player 1 in Kuhn poker during forward pass of the algorithm. Backward pass that uses these values is shown in Table 2. In our representation history $h$ is a concatenation of all public and private actions. The game tree trajectory column shows the path in the game tree that was sampled. Solid arrows denote sampled actions while dashed arrows show other available actions, all actions have their probability under current strategy $\sigma$ next to them. The sampled history in this case is: chance deals (K)ing to player 1, chance deals (Q)ueen to player 2, player 1 (B)ets, player 2 (C)alls. We will use shorter notation $KQBC$ to refer to this history. For each history $h$ reach probability $\pi_{-1}^\sigma(h)$ shows how likely the history is reached when player 1 plays in a way to get to this history. The sampling probabilities $q(h)$ are computed following sampling policy $\xi$ which is uniform in this case, i.e. for each history all available actions have the same probability that they will be sampled. The last two columns show augmented information sets for each player in each history. For example for player 1 history KQB is represented by information set K?B since he does not know what card was dealt to PLAYER 2. Light gray background marks cells where the values are well defined however they are not used in our example update for player 1.
Table 2: The backward pass starts by evaluating utility of the terminal history: $\hat{u}_1^b(\sigma, KQBC|KQBC) = +2$ since player 1 has (K)ing which is better card than opponent’s (Q)ueen. In the next step computation updates values for history KQB. Expected baseline corrected history-action value $\hat{u}_1^b(\sigma, KQB,Call|KQBC)$ is computed based on current sample and then used together with $\hat{u}_1^b(\sigma, KQ, Bet|KQBC)$ to compute $\hat{u}_1^b(\sigma, KQB|KQBC)$. When updating values for history KQB baseline corrected sampled counterfactual values are computed based on just updated $u_1^b(\sigma, KQ, Bet|KQBC)$ for the sampled Bet action and on a baseline value $\hat{u}_1^b(\sigma, KQ, Check|KQBC)$ for Check action that was not sampled. Reach probability $\pi_{\sigma_0}(KQ)$ and sampling probability $q(KQ)$ that are also needed to compute counterfactual-values $\hat{v}_1^b(\sigma, K?, a|KQBC)$ were already computed in the forward pass. The counterfactual values are then used to compute actions’ regrets (Eq. 3) which is not shown in the table. Values in cell with light gray background are not used in computation of $\hat{v}_1^b(\sigma, K?, a|KQBC)$.