AIVAT: A New Variance Reduction Technique for Agent Evaluation in Imperfect Information Games

Neil Burch[1], Martin Schmid[1,2], Matej Moravčík[1,2], Dustin Morill[1], Michael Bowling[1]
[1]Department of Computing Science, University of Alberta
[2]Department of Applied Mathematics, Charles University in Prague
{nburch,mschmid,moravcik,morill,mbowling}@ualberta.ca

Abstract

Evaluating agent performance when outcomes are stochastic and agents use randomized strategies can be challenging when there is limited data available. The variance of sampled outcomes may make the simple approach of Monte Carlo sampling inadequate. This is the case for agents playing heads-up no-limit Texas hold'em poker, where man-machine competitions typically involve multiple days of consistent play by multiple players, but still can (and sometimes did) result in statistically insignificant conclusions. In this paper, we introduce AIVAT, a low variance, provably unbiased value assessment tool that exploits an arbitrary heuristic estimate of state value, as well as the explicit strategy of a subset of the agents. Unlike existing techniques which reduce the variance from chance events, or only consider game ending actions, AIVAT reduces the variance both from choices by nature and by players with a known strategy. The resulting estimator produces results that significantly outperform previous state of the art techniques. It was able to reduce the standard deviation of a Texas hold'em poker man-machine match by 85% and consequently requires 44 times fewer games to draw the same statistical conclusion. AIVAT enabled the first statistically significant AI victory against professional poker players in no-limit hold'em. Furthermore, the technique was powerful enough to produce statistically significant results versus individual players, not just an aggregate pool of the players. We also used AIVAT to analyze a short series of AI vs human poker tournaments, producing statistical significant results with as few as 28 matches.

Introduction

Evaluating an agent's performance in stochastic settings can be hard. Non-zero variance in outcomes means the game must be played multiple times to compute a confidence interval that likely contains the true expected value. Regardless of whether the variance arises from player actions or from chance events, we might need to observe many samples before we get a narrow enough interval to draw desirable conclusions. In many situations, it is simply not feasible (e.g., when the evaluation involves human participation) to simply observe more samples, so we must turn to statistical techniques that use additional information to help narrow the confidence interval.

This agent evaluation problem is commonly encountered in games, where the goal is to estimate the expected performance difference between players. For example, consider poker games. Poker is not only a long-standing challenge problem for AI (von Neumann 1928; Koller and Pfeffer 1997; Billings *et al.* 2002) with annual competitions (Zinkevich and Littman 2006; Annual Computer Poker Competition), but also a very popular game played by an estimated 150 million players worldwide (Eco 2007). Heads-up no-limit Texas hold'em (HUNL) is a particular variant of the game that has received considerable attention in the AI community in recent years, including a "Brains vs. AI" event pitting Claudico (Brains Vs. AI 2015), a top HUNL computer program, against professional poker players. This match involved 80,000 hands of poker with four poker players playing dozens of hours each over thirteen days. Despite Claudico losing by over 9 big blinds per 100 hands (a margin that is considered "huge" by poker professionals (Wood 2015)), the result is only on the edge of statistical significance, making it hard to draw a conclusion from this large investment of human time.

Even more recently, two more man-machine competitions were held in close succession. Over a period of four weeks between November 7th and December 12th 2016, DeepStack played a pool of professional poker players recruited by the International Federation of Poker. Soon after, Libratus played a team of four professional heads-up poker specialists in a HUNL competition held January 11-30, 2017. While both computer agents were victorious versus their respective opponents, these two competitions serve as canonical example of the importance of strong variance reduction.

The DeepStack competition used the techniques described in this paper. The Libratus competition used more basic variance reduction methods. As a result, the DeepStack competition was able to achieve statistically significant results on individual participants with only a few hours of play. In comparison, Libratus required full days from the participants for nearly three straight weeks to achieve a result just for the aggregation of all of the players.

Previous techniques for variance reduction in this setting have used two broad classes of statistical techniques.

Techniques like MIVAT (White and Bowling 2009) use the method of control variates with heuristic value estimates to reduce the variance caused by chance events. The technique of importance sampling over imaginary observations (Bowling *et al.* 2008) takes a different approach, using knowledge of a player strategy to evaluate multiple states given a single observation. Imaginary observations can be used to reduce the variance caused by privately observed chance events, as well as the player's random choice whether to make any actions which would immediately end the game.

Techniques from the two classes can be naively combined, but are not specifically designed to work together for the greatest reduction in variance, and neither of the techniques deal with the variance caused by non-terminal action selection. Because good play in imperfect information games generally requires randomised action selection, ignoring action variance is an major shortcoming. We introduce the action-informed value assessment tool (AIVAT), an unbiased low-variance estimator for imperfect information games which extends the use of control variates to player actions, and makes explicit use of imaginary observations to exploit knowledge of the game structure and player strategies.

Background

This paper focuses on variance reduction when evaluating agents for extensive form games, a class of imperfect information sequential decision making problems. Formally, an extensive form game involves a set of players P and chance player p_c, a set of states S described as a history of actions from the initial state $\varnothing$, a set $Z \subset S$ of terminal states, acting player $p(h) : S \setminus Z \mapsto P \bigcup \{p_c\}$, player value functions $v_p(z) : Z \mapsto \mathbf{R}$, and information partitions $\mathcal{I}_p$ of $\{h \in S | p(h) = p\}$. We will say $h \sqsubset h'$ if a game in state h' was previously in state h, $h \sqsubseteq h'$ if $h \sqsubset h'$ or $h = h'$, $A(h)$ is the set of valid actions at h, and $h \cdot a$ is the successor state of h that is reached by making action a. For all states h such that $p(h) = p_c$, $\sigma_{p_c}(h, a)$ is the publicly known probability distribution over possible chance outcomes at state h.

An information set $I \in \mathcal{I}_p$ describes a set of states that player p can not distinguish due to imperfect information of the game state. Any player decision is therefore made at information sets, not states. A behaviour strategy $\sigma_p(I, a)$ gives the probability of player p making decision a at information set I. The behaviour in a state is determined by the information set I, so that $\forall h \in I \; \sigma_p(h, a) = \sigma_p(I, a)$. We will say the probability of reaching a state h is $\pi(h) = \Pi_{h' \cdot a} \sigma_{p(h')}(h', a)$. It is also useful to consider $\pi_p(h) = \Pi_{h' \cdot a \sqsubseteq h, p(h') = p} \sigma_p(h', a)$, the probability of a player reaching state h if all other players play to reach h. This notation can be extended so that for any set of players T, $\pi_T(h) = \Pi_{p \in T} \pi_p(h)$.

When talking about estimating the value for players in a game, we are trying to find the expected value $\mathbb{E}_z[v_p(z)] = \sum_{z \in Z} \pi(z) v_p(z)$. An estimator $e(z)$ is said to be unbiased if the expected value $\mathbb{E}_z[e(z)] = \mathbb{E}_z[v_p(z)]$. Having an estimator be provably unbiased is important because it is in some sense truthful: a player can not appear to do better by changing their play to take advantage of the estimation method.

MIVAT and Imaginary Observations

AIVAT is an extension of two earlier techniques, MIVAT and importance sampling over imaginary observations. MIVAT (White and Bowling 2009) and its precursor DIVAT (Zinkevich *et al.* 2006) use value functions for a control variate that estimates the expected utility given observed chance events. Conceptually, the techniques subtract the change in expected utility due to a chance action to get a lower variance value. For example, in poker, it is likely that good hands end in positive outcomes and bad hands end in negative outcomes. Starting with the observed outcome, we could subtract some value for good hands and add a value for bad hands, and we would expect the corrected value to have lower variance. If the expected value of the correction terms is zero, we can use the lower variance corrected value as an unbiased estimator of player value.

DIVAT requires a strategy for all players to generate value estimates for states through self-play, while MIVAT generalized the approach by allowing for arbitrary value functions defined after chance events. Both add a correction term for each chance event in an observed state. In order to remain unbiased despite using an arbitrary value estimation function $u(o)$, MIVAT uses a correction term of the form $\mathbb{E}_o[u(o)] - u(o)$ for an observation with outcome o. Computing this expectation requires us to know the probability distribution that o was drawn from, which is true in the case of chance events as σ_{p_c} is public knowledge. These terms are guaranteed to have an expected value of zero, making the MIVAT value (observed value plus correction terms) an unbiased estimate of player value. In a game like poker, MIVAT will account for the dealer giving a player favourable or unfavourable cards, but not for lucky player actions selected from a randomised strategy.

Imaginary observations with importance sampling (Bowling *et al.* 2008) uses knowledge of a player's strategy to compute an expected value of multiple states given an observation of a single state. Due to imperfect information, there may be many states which are all guaranteed to have the same probability of the opponent making their actions. If we consider importance sampling over these imaginary observations, the opponent's probability of reaching the state cancels out so we do not need to know the opponent's strategy. By taking an expectation over a set of states for every observation, we expect a lower variance value.

There are two kinds of situations where we can use imaginary observations. First, for any states h where player p could have made an action a which ends the game, we can add the imaginary observation of the terminal state $h \cdot a$. For example, in poker this lets us consider player p folding to a bet they called or raised, or calling a bet we folded to in the final round. Second, because of the information partitions in imperfect information games, there may be other states that have identical opponent probabilities. In poker, this lets us consider all the states where the public player actions are the same, the opponent private cards and public board cards are the same, but player p has different private cards. Imaginary observations do not let us reduce the variance caused by choosing non-terminal actions or the outcomes of publicly visible chance events.

MIVAT and imaginary observations consider different information and can be combined to get a value estimate with lower variance than either technique used individually. Instead of using the terminal value $v(z)$ for an imaginary observation z, we could use the MIVAT value estimate given z. However, because neither technique has terms which address the effect of non-terminal actions, we would never expect this combination of techniques to produce a zero variance value estimate. Even with a "perfect" value function that correctly estimates the expected value of a state and action for the players, there would still be some variance in the value estimate due to random action selection by players.

AIVAT

Conceptually, AIVAT combines the chance correction terms of MIVAT with imaginary observations across private information, along with new MIVAT-like correction terms for player actions. The AIVAT estimator is the sum of a base value using imaginary observations, plus imaginary observation correction terms for both player actions and chance events. Roughly speaking, moving backwards through the choices in an observed game, the AIVAT correction terms are constructed in a fashion that shifts an estimate of the expected value after a choice was made towards an estimate of the expected value before the choice.

Because imaginary observations with importance sampling provides an unbiased estimate of the expected value of the players, and the MIVAT-like terms have an expected value of zero, AIVAT is also an unbiased estimator of the expected player value. Furthermore, with well-structured games, "perfect" value functions, and knowledge of all player strategies, we could achieve zero variance estimators: the imaginary observation values and the correction terms would sum to the expected player value, regardless of the observed game.

Figure 1 gives a high level overview of MIVAT, imaginary observations, and AIVAT. In this example, we are interested in the expected value for player 1, and know player 1's strategy. We use an observation of one hand of Leduc hold'em poker, a small synthetic game constructed for research (Southey *et al.* 2005). Leduc hold'em is a two round game with one private card for each player, and one publicly visible board card that is revealed after the first round of player actions. In the example, player 1 is dealt **Q♠** and player 2 is dealt **K♠**. Player 1 makes the **check** action followed by a player 2 **check** action. The public board card is revealed to be **J♡**. There are three actions in round two: **check**, **raise**, **call** ; and player 1 loses 5 chips.

AIVAT Correction Terms

We start by describing the correction terms added for chance events and actions. Given information about a player's strategy, we can treat that player's choice events as chance events and construct MIVAT-like correction terms for them. The player strategy also allows imaginary observations considering alternative histories with identical opponent probabilities, so we can compute an expectation over a set of compatible histories rather than using the single observed outcome.

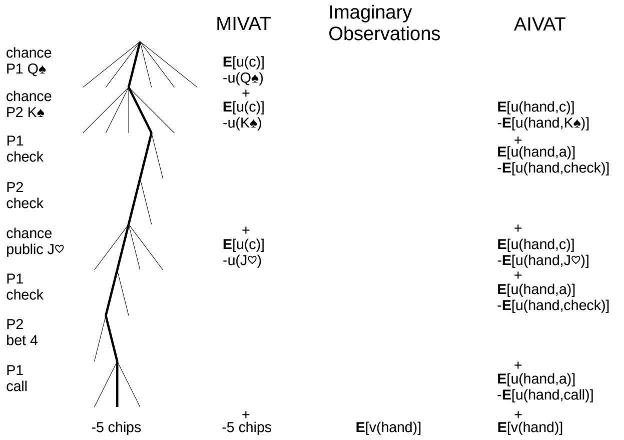

Figure 1: Comparison of MIVAT, imaginary observations, and AIVAT

The correction term at a decision point will be the expectation across all compatible histories of the expected value before a choice, minus the value after the observed choice. As with MIVAT, the values are estimated using an arbitrary fixed value function to estimate the value after every decision. Value estimates which more closely approximate the true expected value will result in greater variance reduction.

To consider imaginary observations, we need at least one player for which we know the strategy. Let P_a be a non-empty set of players, including p_c, such that $\forall p \in P_a$ we know σ_p, and $P_o = P \setminus P_a$ be the set of opponent players for which we do not know the strategy. If $P_a = \{p_c\}$ then AIVAT would be identical to MIVAT. We must also partition the states into the sets we can evaluate given an observation of a completed game. Let $\mathcal{H}$ be a partition of states $\{h | p(h) \in P_a\}$ such that $\forall H \in \mathcal{H}$ and $\forall h, h' \in H$,

1. $\forall p \in P_o \; \forall \sigma_p \; \pi_p(h) = \pi_p(h')$. For example, this can be enforced by requiring h and h' to pass through the same sequence of player p information sets and make the same actions at those information sets.
2. $h \not\sqsubseteq h'$. This implies a uniqueness property, where for any terminal z, $\{h'' | h'' \sqsubset z, h'' \in H\}$ is either empty or a singleton.
3. We will extend the actions so that $A'(h) = \bigcup_{h'' \in H} A(h'')$ and let $\sigma(h, a) = 0 \; \forall a \in A'(h) \setminus A(h)$. Because $A'(h) = A'(h')$ we will say $A(H) = A'(h)$.

Similar to MIVAT, we need value functions that give an estimate of the expected value after an action. Let there be arbitrary functions $u_h(a) : A'(h) \mapsto \mathbf{R}$ for each state h where $p(h) \in P_a$. Say we have seen a terminal state z. Consider a part $H \in \mathcal{H}$. If $\nexists h \in H$ such that $h \sqsubset z$, then the correction term $k_H(z) = 0$. Otherwise, property 2 of $\mathcal{H}$ implies there is a unique observed action a_O such that

$h \cdot a_{\mathrm{O}} \sqsubseteq z, h \in H, a_{\mathrm{O}} \in A(h)$, and the correction term is

$$k_H(z) = \frac{\sum_{a\in A(H)} \sum_{h\in H} \pi_{P_a}(h \cdot a) u_h(a)}{\sum_{h\in H} \pi_{P_a}(h)} - \frac{\sum_{h\in H} \pi_{P_a}(h \cdot a_{\mathrm{O}}) u_h(a_{\mathrm{O}})}{\sum_{h\in H} \pi_{P_a}(h \cdot a_{\mathrm{O}})}$$

AIVAT uses the sum of $k_H(z)$ across all $H \in \mathcal{H}$.

AIVAT Base Value

The AIVAT correction terms have an expected value of zero, and are not a value estimate by themselves. They must be combined with an unbiased estimate of player value. For improved variance reduction, the form of the correction terms must match the choice of base value estimate.

To see how the terms match, consider a simplified version of AIVAT where the final correction term for a terminal state $h \cdot o$ has the form $\mathbb{E}_a[u_h(a)] - u_h(o)$. Ideally, we would like the value estimate for $h \cdot a$ to be $u_h(a)$. The value estimate plus the correction term will then have the same value $\mathbb{E}_a[u_h(a)]$ for all actions at h, resulting in zero variance.

For the AIVAT correction terms, the correct choice is to use imaginary observations of all possible private information for players in P_a, as in "Example 3: Private Information" of the paper by Bowling *et al.* (Bowling *et al.* 2008). In poker, it corresponds to evaluating the game with all possible private cards, weighted by the likelihood of holding the cards given the observed game. For completeness, we formally describe the particular instance of this existing estimator using the notation of this paper.

Given the correction term partition $\mathcal{H}$ of player P_a states, we construct a matching partition $\mathcal{W}$ of terminal states such that $\forall W \in \mathcal{W}$ and $\forall z, z' \in W$,

- $\forall p \in P_o \; \forall \sigma_p \; \pi_p(z) = \pi_p(z')$.
- a player in P_a made an action in $z \iff$ a player in P_a made an action in z'.
- if a player in P_a made an action in z, then for the longest prefix $h \sqsubset z$ and $h' \sqsubset z'$ such that $p(h) \in P_a$ and $p(h') \in P_a$, both h and h' are in the same part of $\mathcal{H}$.

The last two conditions on $\mathcal{W}$ ensure that the imaginary observation estimate does not include terminal states that the correction terms will also account for. This rules out a form of double counting which would not produce a biased estimator, but would increase the variance when using high quality estimates in the correction terms.

If we observe a terminal state z, let $W \in \mathcal{W}$ be the part such that $z \in W$. The base estimated value for player p is

$$\frac{\sum_{z' \in W} \pi_{P_a}(z') v_p(z')}{\sum_{z' \in W} \pi_{P_a}(z')}$$

AIVAT Value Estimate

The AIVAT estimator gives an unbiased estimate of the expected value $\mathbb{E}_z[v_p(z)]$. If we use partitions $\mathcal{H}$ and $\mathcal{W}$ as described above, and are given an observation of a terminal state $z \in W \in \mathcal{W}$, the value estimate is

$$\text{AIVAT}(z) = \frac{\sum_{z' \in W} \pi_{P_a}(z') v_p(z')}{\sum_{z' \in W} \pi_{P_a}(z')} + \sum_{H \in \mathcal{H}} k_H(z) \quad (1)$$

Note that there is a subtle difference between AIVAT and a simple combination of imaginary observations with an extended MIVAT framework using player strategy information to add control variates for actions. Using extended MIVAT plus imaginary observations, we would consider the expected MIVAT value estimate across all terminal histories compatible with the observed terminal state. In AIVAT, for each correction term we would consider all histories compatible with the state at that decision point.

As a concrete example of the difference, consider the game used in Figure 1. MIVAT with imaginary observations would only consider private cards for player 1 that do not conflict with the opponent's **K♠** or the public card **J**♡, even when computing the $\mathbb{E}[u(c)] - u(\mathbf{J}\heartsuit)$ control variate term for the public card. In contrast, AIVAT considers **J**♡ as a possible player card for the term.

Unbiased Value Estimate

It is desirable to have an unbiased value estimate for games, so that players can not improve their estimated value by changing their strategy to fit the estimation technique. We prove that AIVAT is unbiased. The value estimate AIVAT(z) in Equation 1 is a sum of two parts. The fraction in the first part is an unbiased estimator based on imaginary observations (Bowling *et al.* 2008), so we only need to show that the sum of all k_H terms has an expected value of 0.

Lemma 1 $\forall H \in \mathcal{H} \; \mathbb{E}_{z\in Z}[k_H(z)] = 0$

Proof. Consider an arbitrary $H \in \mathcal{H}$. Let $Z(H) = \{z \in Z | \exists h \in H, h \sqsubset z\}$ be the set of terminal states passing through H. Expanding definitions, using property 1 of $\mathcal{H}$ and multiplying by $\pi_{P_o}(H)/\pi_{P_o}(H) = 1$ we get

$$\begin{aligned}
\mathop{\mathbb{E}}_{z\in Z}[k_H(z)] &= \sum_{z\in Z} \pi(z) k_H(z) = \sum_{z \in Z(H)} \pi(z) k_H(z) \\
&= \sum_{z\in Z(H)} \pi(z) \frac{\pi_{P_o}(H)}{\pi_{P_o}(H)} \frac{\sum_{a\in A(H)} \sum_{h\in H} \pi_{P_a}(h \cdot a) u_h(a)}{\sum_{h\in H} \pi_{P_a}(h)} \\
&\quad - \sum_{z\in Z(H)} \pi(z) \frac{\pi_{P_o}(H)}{\pi_{P_o}(H)} \frac{\sum_{h\in H} \pi_{P_a}(h \cdot a_{\mathrm{O}}) u_h(a_{\mathrm{O}})}{\sum_{h\in H} \pi_{P_a}(h \cdot a_{\mathrm{O}})}
\end{aligned}$$

Using $\pi_{P_o}(h)\pi_{P_a}(h) = \pi(h)$

$$\begin{aligned}
&= \sum_{z\in Z(H)} \pi(z) \frac{\sum_{a\in A(H)} \sum_{h\in H} \pi(h \cdot a) u_h(a)}{\sum_{h\in H} \pi(h)} \\
&\quad - \sum_{z\in Z(H)} \pi(z) \frac{\sum_{h\in H} \pi(h \cdot a_{\mathrm{O}}) u_h(a_{\mathrm{O}})}{\sum_{h\in H} \pi(h \cdot a_{\mathrm{O}})}
\end{aligned}$$

Using $\sum_{z,h\sqsubset z}\pi(z)=\pi(h)$ and $\sum_{z,h\cdot a\sqsubset z}\pi(z)=\pi(h\cdot a)$

$$\begin{aligned}&= \sum_{h'\in H}\pi(h')\frac{\sum_{a\in A(H)}\sum_{h\in H}\pi(h\cdot a)u_h(a)}{\sum_{h\in H}\pi(h)}\\ &- \sum_{h'\in H}\sum_{a\in A(h')}\pi(h'\cdot a)\frac{\sum_{h\in H}\pi(h\cdot a)u_h(a)}{\sum_{h\in H}\pi(h\cdot a)}\end{aligned}$$

Using property 3 of $\mathcal{H}$

$$\begin{aligned}&= \sum_{h'\in H}\pi(h')\frac{\sum_{a\in A(H)}\sum_{h\in H}\pi(h\cdot a)u_h(a)}{\sum_{h\in H}\pi(h)}\\ &- \sum_{a\in A(H)}\sum_{h'\in H}\pi(h'\cdot a)\frac{\sum_{h\in H}\pi(h\cdot a)u_h(a)}{\sum_{h\in H}\pi(h\cdot a)}\\ = &\sum_{a\in A(H)}\sum_{h\in H}\pi(h\cdot a)u_h(a) - \sum_{a\in A(H)}\sum_{h\in H}\pi(h\cdot a)u_h(a)\\ &= 0\end{aligned}$$

Because the expected value is 0 for an arbitrary H, the expected value is 0 for the sum of all $H\in\mathcal{H}$. ■

Theorem 1 $\mathbb{E}_{z\in Z}[\sum_{H\in\mathcal{H}} k_H(z)] = 0$

Proof. This immediately follows from Lemma 1, as the expected value of a sum of terms is the sum of the expected values of the terms, which are all 0. ■

Experimental Results

We demonstrate the effectiveness of AIVAT in two poker games, Leduc hold'em and heads-up no-limit Texas hold'em (HUNL). Both Leduc hold'em and HUNL have a convenient structure where all actions are public, and there is a mix of chance events in the form of completely public board cards and completely private hole cards. The uncomplicated structure leads to a clear choice for the partition $\mathcal{H}$. Each $H\in\mathcal{H}$ has states with identical betting, public board cards, and private hole cards for any players in P_o.

For the first experiments where we compare against previous techniques, the value functions $u_h(a)$ are self-play values, generated by solving the game to find a Nash equilibrium strategy using a variant of the Monte Carlo CFR algorithm (Lanctot *et al.* 2009). For each player p_x and partition H, we save the average observed values for opponent p_y across all iterations, giving us a value $w_H(a) \approx \sum_{h\in H}\pi_{p_x}(h\cdot a)\mathbb{E}[v_{p_y}(h)]/\sum_{h\in H}\pi_{p_x}(h\cdot a)$. $w_H(a)$ is an expected self-play value for p_y at H, given the probability distribution of hands for p_x that reach H and play a. Because we are playing a zero-sum game and $v_{p_x}(h) = -v_{p_y}(h)$, we can use $u_h(a) = -w_H(a)\ \forall h\in H$. In HUNL, which is too large to solve directly, we solve a very small abstraction of the game (Billings *et al.* 2003; Ganzfried and Sandholm 2014) with only 8 million information sets, which gives us a rough estimate of $w_H(a)$ that is identical across many partitions of HUNL states.

Poker is played in an alternating fashion, where agents take turns playing in different positions. Let us say we have two agents, x and y. In poker, in odd-numbered games (starting at game 1) we would have x as player 1 and y as player 2, and in even-numbered games we would have y as player 1 and x as player 2. For the experiments, we model this as an extended game where there is an initial 50/50 chance event that assigns a position to the agent, along with an AIVAT correction term for the position.

All experiments will compare AIVAT value estimates with the unmodified game values from counting chips, the MIVAT value estimate, and the combination of MIVAT and imaginary observations using the strategy for agent x (MIVAT+IO$_x$). Because poker is a zero-sum game, it is sufficient to present results from the point of view of agent x.

Leduc Hold'em

The small size of Leduc hold'em lets us test both the case where P_a only contains one non-chance player, as well as the full-knowledge case where $P_a = P$. AIVAT and chip count results are generated from observations of 100,000 games. All of the numbers are in units of chips, where Leduc hold'em has a 1 chip ante, and 2 chip and 4 chip bets in the first and second rounds, respectively.

Table 1 looks at self-play, where both x and y play the same Nash equilibrium that was used to generate $u_h(a)$. The true expected value for player x is 0. Because we are using value functions computed from their self-play, this experiment represents a best-case situation. With knowledge of both player's strategies, the only remaining variance comes from noise in the $u_h(a)$ value function that arises from the sampling and averaging used in the MCCFR computation.

Estimator	$\bar{v}_x$	$SD(v_x)$
chips	0.01374	3.513
MIVAT	0.00448	2.327
MIVAT+IO$_x$	0.00987	1.928
$P_a = \{p_c, x\}$	-0.00009	0.00643
$P_a = \{p_c, x, y\}$	-0.00001	0.00377

Table 1: Value estimates for self-play in Leduc hold'em

With knowledge of both player's strategies, we reduce the per-game standard deviation of the estimated player value by a little less than 99.9%. This situation might be unlikely in practice, but does demonstrate that the AIVAT computation correctly shifts every observed outcome to the expected player value, given full correct information. Surprisingly, the one-sided evaluation where we use only one player's strategy still reduces the standard deviation by 99.8%. Using MIVAT or MIVAT+IO$_x$, we only see a 33.8% and 45.1% reduction, respectively.

Moving away from the best-case situation, Figure 2 looks at games where x is the same Nash equilibrium from above, and y is an agent that randomly calls or raises. Given these strategies, the true expected value for player x is 0.69358.

Using the call/raise strategy for y demonstrates that the amount of variance reduction does depend on how well the value functions estimate the true expected value of a situation. We used value functions which encode self-play values for x, and while y is sufficiently similar to x that the true

Estimator	$\bar{v}_x$	$SD(v_x)$
chips	0.71673	5.761
MIVAT	0.68932	4.412
MIVAT+IO$_x$	0.69968	4.295
$P_a = \{p_c, x\}$	0.69050	1.437
$P_a = \{p_c, x, y\}$	0.68698	1.782
$P_a = \{p_c, y\}$	0.69614	2.983

Table 2: Value estimates for dissimilar strategies in Leduc hold'em

values are still positively correlated with the estimated values for both players, they are no longer an almost-perfect match. Despite the strategic mismatch, using AIVAT we see a reduction in the standard deviation of 48% to 75% compared to the basic chip-count estimate. All of the AIVAT estimators outperform the 25% reduction using MIVAT plus imaginary observations.

No-limit Texas Hold'em with Bots

The game of HUNL better represents a potential real-world application. The game is commonly played, it is too large to easily compute exact expected values directly even when the strategy of both agents is known, average win rate is a statistic of interest to players and observers, and the high per-game variance of outcomes obscures the win rate even after hundreds of thousands of hands.

First, we used data generated by small abstraction-based CFR agents to compare AIVAT to prior techniques. The next section of this paper presents our main results - AIVAT evaluation on human experimental data in this challenging game.

Comparison to previous techniques The variant of HUNL we used here is know as the "Doyle's game" (to be briefly described in the next section). Due to the large branching factor of chance events, we can only present results for AIVAT analysis using the strategy of one agent.

First, we look at self-play using a low-quality Nash equilibrium approximation for both players x and y. The value functions $u_h(a)$ come from this low-quiliaty strategy. Table 3 compares the different techniques.

Estimator	$\bar{v}_x$	$SD(v_x)$
chips	0.03871	25.962
MIVAT	0.02038	21.293
MIVAT+IO$_x$	0.02596	16.073
$P_a = \{p_c, x\}$	0.00186	8.095

Table 3: Selfplay value estimates for self-play in HUNL using 1 million data points. Note that the true expected value for x is 0 since this is in self-play.

In Table 4 we look at games where x uses the same low-quality approximation of a Nash equilibrium, and y is a much stronger agent using a high-quality approximation of a Nash equilibrium. The value functions $u_h(a)$ are still generated using the low-quality approximation. The true expected value for player x is not known.

Estimator	$\bar{v}_x$	$SD(v_x)$
chips	-0.10017	26.308
MIVAT	-0.11565	21.546
MIVAT+IO$_x$	-0.11297	16.051
$P_a = \{p_c, x\}$	-0.10971	8.301

Table 4: Value estimates for dissimilar strategies in HUNL using 1 million data points.

In both experiments, we see a 39% reduction in the standard deviation when using MIVAT with imaginary observations, and a bit more than a 68% reduction using AIVAT. It must be noted that our value function could be improved, as the 18% reduction for MIVAT in this experiment does not match the 23% improvement previously demonstrated using values learned from data (White and Bowling 2009). The small abstract game used to generate the value functions does not do a good job of understanding the consequences of cards being dealt, as it cannot distinguish most card situations. Despite this handicap, the full AIVAT estimator still significantly improves on the state of the art for low-variance value estimators for imperfect information games

No-limit Texas Hold'em with Humans

We used AIVAT variance reduction to evaluate the gameplay of DeepStack, the first poker AI to beat professional human players in no-limit poker. DeepStack recently played two man-machine events with different game rules.

The first format, referred to as "Doyle's Game" is a popular variant amongst the computer poker community — all the previously held man machine competitions (Brains Vs. AI 2015) used this particular game format. This format was used during the DeepStack evaluation: 33 players from 17 different countries were recruited by the International Federation of Poker. The players were asked to complete $3,000$ games each, producing a total of $44,852$ games (not all the players finished the required number of games).

The second format used for evaluation was series of exhibition matches in the "freezeout" poker tournament format, where chips are not reset after each game, and a match continues until one player is out of chips. While much more challenging for AI, this format is very popular by players.

Doyle's Game

In Doyle's Game, the number of players' chips gets reset after every single game, and the goal is to maximize the total amount of chips won in the course of multiple games. Thus, the estimate we are interested in is the expected number of chips won per game.

Based on the aggregate results, DeepStack won the challenge statistically significantly by 4 sigma even with no variance reduction. AIVAT analysis was able to improve the statistical analysis, reducing the standard deviation by 85% and producing 20 sigma significance. Even more impressively, AIVAT resulted in a statistically significant outcome for all but one of the individuals who finished the required number of hands. Concretely, AIVAT estimated that DeepStack was ahead of all 11 players that completed the challenge, and these individual victories were 2 sigma significant for all but one of them. More details are presented in Table 5.

Note that the 85% standard deviation reduction achieved by DeepStack value estimates is substantially lower than the 68% reduction using estimates from the CFR abstraction-based agent. This suggests that DeepStack has a better understanding of the game values compared to the abstraction based approaches.

Table 5: Results against professional poker players estimated with AIVAT (Luck Adjusted Win Rate) and chips won (Unadjusted Win Rate), both measured in mbb/g. Note that 10mbb/g equals 1bb/100. Each estimate is followed by a 95% confidence interval. The double line separates the players that did not finish the required number of $3,000$ hands. Note that for most of the players that finished the match and even for many players that played too few games to finish, the AIVAT produces per-player statistically significant results. This contrasts to other man-machine competitions, where only the aggregate data from all players had to be used to get significant results.

Player	Luck Adjusted Win Rate	Unadjusted Win Rate
Martin Sturc	70 ± 119	-515 ± 575
Stanislav Voloshin	126 ± 103	-65 ± 648
Prakshat Shrimankar	139 ± 97	174 ± 667
Ivan Shabalin	170 ± 99	153 ± 633
Lucas Schaumann	207 ± 87	160 ± 576
Phil Laak	212 ± 143	774 ± 677
Kaishi Sun	363 ± 116	5 ± 729
Dmitry Lesnoy	411 ± 138	-87 ± 753
Antonio Parlavecchio	618 ± 212	1096 ± 962
Muskan Sethi	1009 ± 184	2144 ± 1019
Pol Dmit	1008 ± 156	883 ± 793
Tsuneaki Takeda	627 ± 231	-333 ± 1228
Youwei Qin	1306 ± 331	1953 ± 1799
Fintan Gavin	635 ± 278	-26 ± 1647
Giedrius Talacka	1063 ± 338	459 ± 1707
Juergen Bachmann	527 ± 198	1769 ± 1662
Sergey Indenok	881 ± 371	253 ± 2507
Sebastian Schwab	1086 ± 598	1800 ± 2162
Dara O'Kearney	78 ± 250	223 ± 1688
Roman Shaposhnikov	131 ± 305	-898 ± 2153
Shai Zurr	499 ± 360	1154 ± 2206
Luca Moschitta	444 ± 580	1438 ± 2388
Stas Tishekvich	-45 ± 433	-346 ± 2264
Eyal Eshkar	18 ± 608	715 ± 4227
Jefri Islam	997 ± 700	3822 ± 4834
Fan Sun	531 ± 774	-1291 ± 5456
Igor Naumenko	-137 ± 638	851 ± 1536
Silvio Pizzarello	1500 ± 2100	5134 ± 6766
Gaia Freire	369 ± 136	138 ± 694
Alexander Bös	487 ± 756	1 ± 2628
Victor Santos	475 ± 462	-1759 ± 2571
Mike Phan	-1019 ± 2352	-11223 ± 18235
Juan Manuel Pastor	2744 ± 3521	7286 ± 9856
Human Professionals	486 ± 40	492 ± 220

Freezouts

In a freezout tournament, players' chips don't reset after every hand, but the game continues until one of the players loses all of their chips, thus losing the tournament. Consequently, a tournament is a single Bernoulli trial and we are interested in an estimate of the win probability (in contrast to chips won per hand in Doyle's game).

It turns out that the AIVAT technique can readily be applied in these settings as well. We simply use a value estimate corresponding to an estimate of the win probability. We opted for the ICM formulae (Ganzfried and Sandholm 2008) to estimate win probability from the players' chips since it has been shown to be very accurate for heads-up (Miltersen and Sørensen 2007). In the case of two players p_1 and p_2 it simplifies down to:

$$ICM(chips_{p_1}, chips_{p_2}) = \frac{chips_{p_1}}{chips_{p_1} + chips_{p_2}} \quad (2)$$

Note that since we use this formula only for the value estimate for the correction terms, the formula can be arbitrary and still be unbiased. But a better estimate of the real win probability leads to lower variance.

Since DeepStack does not use pre-computed strategy but rather computes the strategy online, it is not limited to a fixed amount of chips and thus can readily play freezout games. DeepStack was used for another experiment versus human poker players, this time in a freezout setting and AIVAT was used to reduce the variance. The resulting values are in the Table 6 and AIVAT was able to produce statistically significant results with as little as 28 freezout matches, while the baseline estimate from 14 wins and 14 losses provided no separation in the estimated skill.

Estimator	95% conf. interval
Bernoulli	0.5 ± 0.2
AIVAT	0.59 ± 0.018

Table 6: Freezout results.

Conclusions

We introduced a technique for value estimation in imperfect information games that extends and combines existing techniques. AIVAT uses heuristic value functions, knowledge of game structure, and knowledge about player strategies to both add a control variate term for chance and player decisions, and to average over multiple possible outcomes given a single observation. We prove AIVAT is unbiased, and demonstrate that with (almost) perfect value functions we see (almost) complete elimination of variance. Even with imprecise value functions, we show variance reduction in a real-world game that significantly exceeds existing techniques. AIVAT's three times reduction in standard deviation allows us to achieve the same statistical significance with ten times less data.

A factor of forty is substantial: for problems with limited data, like human play against bots, forty times as many games could be the distinction between practical and impractical experiments.

References

Annual computer poker competition. website, http://www.computerpokercompetition.org/.

Darse Billings, Aaaron Davidson, Jonathen Schaeffer, and Duane Szafron. The challenge of poker. *Artificial Intelligence*, 134(1–2):201–240, 2002.

Darse Billings, Neil Burch, Aaaron Davidson, Robert Holte, Jonathan Schaeffer, Terence Schauenberg, and Duane Szafron. Approximating game-theoretic optimal strategies for full-scale poker. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 661–668, 2003.

Michael Bowling, Michael Johanson, Neil Burch, and Duane Szafron. Strategy evaluation in extensive games with importance sampling. In *Proceedings of the Twenty-Fifth International Conference on Machine Learning (ICML)*, pages 72–79, 2008.

Brains Vs. AI. http://www.cs.cmu.edu/brains-vs-ai, 2015.

Poker: A big deal. *The Economist*, December 22:31–38, 2007.

Sam Ganzfried and Tuomas Sandholm. Computing an approximate jam/fold equilibrium for 3-player no-limit texas hold'em tournaments. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*, pages 919–925. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

Sam Ganzfried and Tuomas Sandholm. Potential-aware imperfect-recall abstraction with earth mover's distance in imperfect-information games. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 682–690, 2014.

Daphne Koller and Avi Pfeffer. Representations and solutions for game-theoretic problems. *Artificial Intelligence*, 94:167–215, 1997.

Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. Monte Carlo sampling for regret minimization in extensive games. In *Advances in Neural Information Processing Systems 22 (NIPS)*, pages 1078–1086, 2009.

Peter Bro Miltersen and Troels Bjerre Sørensen. A near-optimal strategy for a heads-up no-limit texas hold'em poker tournament. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, page 191. ACM, 2007.

Finnegan Southey, Michael H. Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and D. Chris Rayner. Bayes' bluff: Opponent modelling in poker. In *UAI '05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence*, pages 550–558, 2005.

J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

Martha White and Michael H. Bowling. Learning a value analysis tool for agent evaluation. In *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, 2009*, pages 1976–1981, 2009.

Jocelyn Wood. Doug polk and team beat claudico to win $100,000 from microsoft & the rivers casino. *Pokerfuse,* http://pokerfuse.com/news/media-and-software/26854-doug-polk-and-team-beat-claudico-win-100000-microsoft/, 2015.

Martin Zinkevich and Michael Littman. The AAAI computer poker competition. *Journal of the International Computer Games Association*, 29, 2006. News item.

Martin Zinkevich, Michael H. Bowling, Nolan Bard, Morgan Kan, and Darse Billings. Optimal unbiased estimators for evaluating agent performance. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, pages 573–579, 2006.