Optimal Unbiased Estimators for Evaluating Agent Performance

Martin Zinkevich and **Michael Bowling** and **Nolan Bard** and **Morgan Kan** and **Darse Billings**
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada T6G 2E8
{maz,bowling,nolan,mkan,darse}@cs.ualberta.ca

Abstract

Evaluating the performance of an agent or group of agents can be, by itself, a very challenging problem. The stochastic nature of the environment plus the stochastic nature of agents' decisions can result in estimates with intractably large variances. This paper examines the problem of finding low variance estimates of agent performance. In particular, we assume that some agent-environment dynamics are known, such as the random outcome of drawing a card or rolling a die. Other dynamics are unknown, such as the reasoning of a human or other black-box agent. Using the known dynamics, we describe the complete set of all unbiased estimators, that is, for any possible unknown dynamics the estimate's expectation is always the agent's expected utility. Then, given a belief about the unknown dynamics, we identify the unbiased estimator with minimum variance. If the belief is correct our estimate is optimal, and if the belief is wrong it is at least unbiased. Finally, we apply our unbiased estimator to the game of poker, demonstrating dramatically reduced variance and faster evaluation.

Introduction

Poker is a game of both skill and chance. As a result, it can be difficult to distinguish between the effects of skill and chance on one's winnings, possibly resulting in disastrous losses. If each player actually received their expected value each hand, it would readily become apparent to a losing player that they should change strategies or stop playing.

Stochastic environments, which combine chance and skill are pervasive in artificial intelligence. However, AI researchers face the same problem that poker players do: it is difficult even after a match is over to evaluate a player or algorithm's performance. The usual solution is repeated independent trials. If two stationary poker algorithms are being compared, then a very large number of hands can be played and averaged to construct a low variance estimate. When analyzing the performance of a computer program playing against a human, the required number of hands to draw a valid conclusion is simply impractical. In domains where a single round of evaluation is expensive or time-consuming (e.g., TAC (Stone & Greenwald 2005) and RoboCup (Kitano *et al.* 1997)) even program comparisons may require an impractical number of rounds.

Two illustrative techniques have been used in the world of games to address this evaluation problem. The first is exemplified by duplicate bridge, a game played by four or more pairs (teams) of players. A set of boards, or deals of the cards, are generated randomly and each North-South pairing plays all boards once in the North-South position, while rotating to face all possible East-West opponents. The total North-South pairing's score is then compared to all, and only, the other North-South pairings. The pairings being compared have all effectively been dealt the same cards and played the same opponents. Therefore, the luck due to the innate value of being dealt a particular hand is reduced, as well as the variance in the score differences. The problem is that it requires restructuring the game so that multiple pairings can see the same opponents and situations. In addition, a pairing is not evaluated against its actual opponents, but against pairings playing the same opponents. Although computer programs can be replicated to play in both seats, humans are not so easily cloned, nor can they be reliably made to forget previous games when playing new ones with symmetric situations. Lastly, this approach only removes one portion of luck. In poker and other domains, stochastic events affect more than just the initial situation.

A second method is an intellectual poker exercise where a player's performance is compared to how well that player would have done had she known her opponents' cards. This is essentially Sklansky's Fundamental Theorem of Poker (Sklansky 1992). However, this metric is unrealistic in that good poker players will never completely reveal what cards they hold by their actions. Moreover, the technique is *biased* in the sense that one will always do better in expectation if the other player's cards were face up. For some games, low variance unbiased estimators exist (Wolfe 2002), but not in general.

This paper focuses on designing an unbiased estimator for the expected utility of an agent or agents interacting in any stochastic environment. As we have already discussed, the simplest unbiased estimator is just the utility of the agent. However, we will show examples of estimators with lower variance. In particular, we will show how any value function from histories to real numbers can be used to form the basis for an unbiased estimator. The value function can be thought of as a guess of the agent's expected utility for each history. We then show that if the value function is a perfect guess, our technique results in the unbiased estimator with minimum variance. We also show that "similar" value functions have similarly low variance. We conclude with experimental results of how this technique dramatically reduces variance in the game of poker.

Example: Poker

The theoretical results in this paper are broadly applicable to both multiagent and single agent settings. Our empirical results will focus on the game of poker and so we will use it as a motivating example.

Texas Hold'em

There are many variants of poker. Our results focus on **Texas Hold'em**, particularly the two-player limit game. A single hand is played with a shuffled 52-card deck, and consists of four rounds. On the first round (the **pre-flop**), each player receives two private cards. On subsequent rounds, public board cards are revealed (three on the **flop**, one on the **turn**, and one on the **river**). After each of these chance events there is a round of betting, where the players alternately decide to **fold**, **call**, or **raise**.[1] When a player folds, the game ends and the other player wins the **pot**, without revealing their cards. When a player calls, an amount matching the other player's wager is placed into the pot. When a player raises, they match the other players' wager and then put in an additional fixed amount. The players alternate until a player folds, ending the hand, or a player calls, continuing the game to the next round.

There is a limit of four raises per round, so the betting sequence has a finite length. The fixed raise amount in the first two rounds is called the **small bet**, which is doubled (a **big bet**) in the last two rounds. If neither player folds before the final betting round is over, there is a **showdown**. The players reveal their private cards and the player who can make the strongest five-card poker hand using any combination of their private cards and the public cards wins the pot. The pot is split in the case of a tie.

Luck and Skill

Consider the following hand of limit Texas Hold'em between Alice and Bob. Alice is dealt **J♠J♣**, and Bob is dealt **6♡5♡**. Alice raises, and Bob calls. Then three cards (the *flop*) are placed on the board, **6♠6♢3♢**. Alice bets, and Bob calls. In the next round, the **T♣** arrives, Alice bets, Bob raises, and Alice calls. Next the **8♡** is dealt, Alice checks, Bob bets, and Alice calls. In the showdown, Bob wins, with three sixes beating two pair.

Consider how much an outside observer might expect Alice to win on a *typical hand* given what happened on this hand. One naive assessment is to focus on the final outcome and conclude that Bob winning nine small bets from Alice is typical. This conclusion ignores the fact that the outcome is decided by more than just the players' decisions – luck plays a large role. One could instead examine the player's decisions alone. In the first round, Alice has a large advantage. If Bob could see Alice's cards, he would fold, since that would lose less in expected value. However, his call is certainly not a bad play in general, and he only lost as much as one is expected to lose in that situation. Bob then got a lucky flop to make a very strong hand. By not raising Alice's bet, he lost a sizable fraction of a bet, but this may have been a **trap** – a deliberate deception to gain more later when the bet size doubles. The turn is rather uninteresting, in that Alice lost only as much as one would expect to lose with her strong hand. However, her check and call (as opposed to the typical bet and call) on the river was insightful, losing one big bet less than would normally be expected. Overall, Alice should be considered to have outplayed Bob on this hand, despite losing a substantial pot, which was the result of an unlucky flop.

Of course, there is the question of how to assign numerical values to each of the players' decisions. We will also want to do so in a way that is unbiased, so we still are estimating the true value of the game. In the next section, we will introduce a formalism that will help us construct unbiased estimators of a game's outcome.

Formalism

Before delving deep into the notation, definitions, and theoretical results, we begin with a high-level overview of the next two sections. Our goal is to construct a low variance estimator for an agent or agents' performance. We assume that certain aspects of the domain or agent are *known*. In addition we have a *belief* or *guess* about all aspects of the system. We construct an estimator that is provably unbiased for any domain consistent with our knowledge. We go on to show that if our guess is (nearly) accurate, the estimator has (nearly) the minimum variance of all unbiased estimators.

Formally, define the set of all atomic events, either actions or chance happenings, as the event set E. We define the sequence of all events that have occurred so far to be the history $h \in E^*$. Define $H \subseteq E^*$ to be the set of all reachable histories, and $O \subseteq H$ to be the set of terminal histories, or **outcomes**. Let us suppose that there is a **utility function** $u : O \to \mathbf{R}$ associating every outcome with a utility. This could represent points, money earned, or a $1, \frac{1}{2}, 0$ value indicating win, tie, and loss respectively.[2]

In this work, we will think about the probability of the next event given a sequence of previous events. For all $h \in H \backslash O$, there is an actual distribution $\sigma : H \backslash O \to \Delta(E)$ over the next event in the sequence, and we will write $\sigma(e|h)$ to be the probability that e is the next event given history h. Now suppose that some of the game or system's dynamics are known. So, there exists a set $K \subseteq H \backslash O$, such that a distribution $k : K \to \Delta(E)$ over the next event in the sequence is known. We will write $k(e|h)$ to be the known probability of e being the next event given h. Note that we have chosen notation such that we can represent the case where the randomness in the system is known and the agents' behavior is unknown (e.g., humans playing poker) and we can

[1] A call or raise when there is no wager by the opponent to match is called a **check** or **bet**, respectively.

[2] In fact, this "utility" could be any real-valued function of the outcome of the game, even if it was not a metric of performance.

represent the case where the environment is unknown and the agents' behavior is known (e.g., a robot in an unknown environment), or some mixture (e.g., a robot and a human playing poker).

Define $\mathcal{K} = \Delta(E)^K$ to be the set of all k functions, and $\Sigma = \Delta(E)^{H \setminus O}$ to be the set of all σ functions. We will say that $k \in \mathcal{K}$ and $\sigma \in \Sigma$ **agree** if for all $h \in K$, $k(h) = \sigma(h)$. Define Σ_k to be the set of all σ that agree with k. Lastly, define $|h|$ to be the number of events in the sequence h, h_i to be the ith event in h, and $h(i)$ to be the first i events of h.

Probability, Expectation and Variance

Before discussing performance estimators, we briefly describe the concepts of variance, expectation, and probability. For all $h \in H$, the **probability of $\boldsymbol{h}$ under $\boldsymbol{\sigma}$** is:

$$\mathrm{Pr}_\sigma[h] = \Pi_{t=1}^{|h|} \sigma(h_t | h(t-1)) \tag{1}$$

where $\sigma(h_t|h(t-1))$ is the probability of the tth element of h given the first $t-1$ elements of h. For simplicity, in this paper we will assume O is finite (or equivalently that the game terminates before some number of events T occur). Therefore, for $\sigma \in \Sigma$ and a random variable $f : O \to \mathbf{R}$, the **expected value of $\boldsymbol{f}$ under $\boldsymbol{\sigma}$** is:

$$\mathbf{E}_\sigma[f] = \mathbf{E}_{h \in \sigma}[f(h)] = \sum_{o \in O} \mathrm{Pr}_\sigma[o] f(o) \tag{2}$$

The **variance of $\boldsymbol{f}$ under $\boldsymbol{\sigma}$** is:

$$\mathbf{Var}_\sigma[f] = \mathbf{E}_\sigma[f^2] - (\mathbf{E}_\sigma[f])^2 \tag{3}$$

For $h, h' \in H$, we'll say $h \sqsubseteq h'$ if h is a prefix of h', or formally $h = h'(|h|)$. Then, if $\mathrm{Pr}_\sigma[h] > 0$, the **conditional probability of $\boldsymbol{h'}$ given $\boldsymbol{h}$ under $\boldsymbol{\sigma}$** and the **conditional expectation of $\boldsymbol{f}$ given $\boldsymbol{h}$ under $\boldsymbol{\sigma}$** are:

$$\mathrm{Pr}_\sigma[h'|h] = I(h \sqsubseteq h') \frac{\mathrm{Pr}_\sigma[h']}{\mathrm{Pr}_\sigma[h]} \tag{4}$$

$$\mathbf{E}_\sigma[f|h] = \sum_{h' \in O} f(h') \, \mathrm{Pr}_\sigma[h'|h] \tag{5}$$

where $I(\text{true}) = 1$ and $I(\text{false}) = 0$. Finally **$\boldsymbol{h}$ is possible under $\boldsymbol{\sigma}$** if $\mathrm{Pr}_\sigma(h) > 0$, **$\boldsymbol{h}$ is possible under $\boldsymbol{k}$** if there is a $\sigma \in \Sigma_k$ where h is possible under σ.

Unbiased Estimators

The goal in this paper is to find performance metrics that are unbiased estimators. Formally, given random variables $\hat{u} : O \to \mathbf{R}$ and $u : O \to \mathbf{R}$:

1. For $\sigma \in \Sigma$, $\hat{u}$ is an **unbiased estimator of $\boldsymbol{u}$ under $\boldsymbol{\sigma}$** if $\mathbf{E}_\sigma[\hat{u}] = \mathbf{E}_\sigma[u]$.
2. For $\Sigma' \subseteq \Sigma$, $\hat{u}$ is an **unbiased estimator of $\boldsymbol{u}$ for $\boldsymbol{\Sigma'}$** if for all $\sigma \in \Sigma'$, $\hat{u}$ is an unbiased estimator of u under σ.
3. $\hat{u}$ is an **unbiased estimator of $\boldsymbol{u}$ for $\boldsymbol{k}$** if $\hat{u}$ is an unbiased estimator of u for Σ_k.

Thus, $\hat{u}$ is an unbiased estimator if, given what we know, regardless of rest of the dynamics, it has the same expected value as u.

In what follows, we will show how to generate an unbiased estimator of u from an informed guess of the expected value of u given h. Up until this point, we have referred to our knowledge k and the true dynamics σ. As suggested by (Harsanyi 1967), instead of considering a situation of incomplete information, we will consider the case where we have imperfect information. In other words we will also consider our **beliefs** about what will happen in any given situation. A belief has the same form as the true dynamics, i.e., it is a function in Σ which may or may not be equal to the true dynamics. However, we will also insist that our beliefs agree (in the formal sense) with our knowledge.

In our development of unbiased estimators we will make use of the concept of a value function $V : H \to \mathbf{R}$. The value $V(h)$ will be thought of as an estimate of the conditional expectation of u given h. Although we will consider all possible value functions in the definitions and main theorem, one natural value function can be derived from our belief about the dynamics. Given our belief $\rho \in \Sigma$ define,

$$V^\rho(h) = \mathbf{E}_\rho[u|h] \tag{6}$$

We will show that with any value function we can generate an unbiased estimator. In addition, we show that a value function from an accurate belief ρ will generate an unbiased estimator with low variance.

We can now describe our proposed estimator. Given $k \in \mathcal{K}$, and a function $V : H \to \mathbf{R}$, define $Q_{V,k} : K \to \mathbf{R}$ such that:

$$Q_{V,k}(h) = \sum_{e \in E} V(h \circ e) k(e|h) \tag{7}$$

where $h \circ e$ is the sequence where e is appended to h. Therefore, $Q_{V,k}$ is a one-step lookahead of the value function given our knowledge. Now define the **advantage sum** $\hat{u}_{V,k} : O \to \mathbf{R}$ to be:

$$\hat{u}_{V,k}(h) = u(h) + \sum_{t \text{ s.t. } h(t) \in K} \left(Q_{V,k}(h(t)) - V(h(t+1)) \right) \tag{8}$$

We replace the effect of every *known random event* on the value of u with the known expected effect of that event.[3]

Theoretical Results

In this paper, we give two sets of theoretical results. The first gives a characterization of the set of unbiased estimators for some given knowledge of the system, which we present in Theorems 1 and 2. The second establishes how to construct unbiased estimators with low variance, which we present as Theorems 3 and 5.

Characterization of Unbiased Estimators

First, we show that a value function can form the basis for an unbiased estimator.

[3]We use the term *advantage sum* to emphasize the similarity to *advantages* in reinforcement learning, which have been shown to be useful in measuring the suboptimality of a policy (Kakade 2003). This work generalizes the idea beyond the knowledge and beliefs commonly used in reinforcement learning, as well as going on to analyze the resulting variance reduction.

Theorem 1 *For any $V : H \to \mathbf{R}$ and $k \in \mathcal{K}$, $\hat{u}_{V,k}$ is an unbiased estimator of u for k.*

Proof: Given $\sigma \in \Sigma_k$. We will prove that every addend in the advantage sum has an expected value of zero. By adding *noop* events, without loss of generality, assume that for all $h \in O$, $|h| = T$ for some T. By linearity of expectation:

$$\mathbf{E}_{h\in\sigma}\left[u(h) + \sum_{t \text{ s.t. } h(t)\in K} (Q_{V,k}(h(t)) - V(h(t+1)))\right]$$

$$= \mathbf{E}_{h\in\sigma}[u] + \sum_{t=1}^{T} \mathbf{E}_{h\in\sigma}\left[\begin{array}{l} I(h(t)\in K) \\ \quad (Q_{V,k}(h(t)) - V(h(t+1))) \end{array}\right]$$

Focusing on a particular summation element t:

$$\begin{aligned} &\mathbf{E}_{h\in\sigma}[I(h(t)\in K)\,(Q_{V,k}(h(t)) - V(h(t+1)))] \\ &= \sum_{h'\in K} \mathbf{E}_{h\in\sigma}[I(h(t)=h')\,(Q_{V,k}(h(t)) - V(h(t+1)))] \end{aligned}$$

Focusing on a particular summation element t and $h' \in K$:

$$\begin{aligned} &\mathbf{E}_{h\in\sigma}[I(h(t)=h')\,(Q_{V,k}(h(t)) - V(h(t+1)))] \\ &= \sum_{e\in E} \mathbf{E}_{h\in\sigma}\left[\begin{array}{l} I(h(t+1)=h'\circ e) \\ \quad (Q_{V,k}(h') - V(h'\circ e)) \end{array}\right] \\ &= \sum_{e\in E} \Pr_\sigma[h'\circ e]\,(Q_{V,k}(h') - V(h'\circ e)) \\ &= \sum_{e\in E} \Pr_\sigma[h']k(e|h')\,(Q_{V,k}(h') - V(h'\circ e)) \qquad (9) \end{aligned}$$

Where Equation 9 follows from the fact that σ and k agree. Since $\sum_{e\in E} k(e|h') = 1$:

$$\begin{aligned} &\mathbf{E}_{h\in\sigma}[I(h(t)=h')\,(Q_{V,k}(h(t)) - V(h(t+1)))] \\ &\qquad = \Pr_\sigma[h']\left(Q_{V,k}(h') - \sum_{e\in E} k(e|h')V(h'\circ e)\right) \end{aligned}$$

By the definition of $Q_{V,k}(h')$, the right side is zero. Therefore, the summation is in expectation zero, implying $\hat{u}_{V,k}$ is an unbiased estimator of u. ■

Moreover, we can characterize any unbiased estimator with a value function.

Theorem 2 *Given any unbiased estimator $\hat{u}$, there is a $V : H \to \mathbf{R}$, such that for all $h \in O$ possible under k, $\hat{u}(h) = \hat{u}_{V,k}(h)$.*

Proof Sketch: We prove the remainder of the theorems in a separate technical report (Zinkevich *et al.* 2006) and merely sketch the reasoning here. The basic argument is that for any unbiased estimator, for any history $h \in H$ possible under k, there is a particular bias for that h, which is independent of the unknown dynamics. Formally, for any $\sigma, \sigma' \in \Sigma_k$ such that $\Pr_\sigma[h] > 0$ and $\Pr_{\sigma'}[h] > 0$:

$$\mathbf{E}_\sigma[\hat{u} - u|h] = \mathbf{E}_{\sigma'}[\hat{u} - u|h] \qquad (10)$$

We then use these biases and some of their basic properties to calculate the value function. ■

Unbiased Estimators of Low Variance

In the previous section we considered the case where we have knowledge of the dynamics of the system, k. We may also have a belief, ρ, about the complete dynamics of the system, which is in agreement with k. We can show that if our belief is correct, i.e., ρ is the same as the true dynamics σ, we can construct a minimum variance unbiased estimator. Formally, given $k \in \mathcal{K}$, $\sigma \in \Sigma_k$, $\hat{u}^*$ is a **minimum variance unbiased estimator for k under σ** if $\hat{u}^*$ is an unbiased estimator for k and for any unbiased estimator for k, $\hat{u}$:

$$\mathbf{Var}_\sigma[\hat{u}^*] \le \mathbf{Var}_\sigma[\hat{u}] \qquad (11)$$

Theorem 3 *For any $k \in \mathcal{K}$, any $\sigma \in \Sigma_k$, $\hat{u}_{V^\sigma,k}$ is a minimum variance unbiased estimator for k under σ.*

Proof Sketch: The first part of the argument involves a non-constructive proof that an unbiased estimator of minimum variance exists. Once this is done, we can prove locally that, for any h possible under k, regardless of the value of V on the remainder of H, having $V(h') = V^\sigma(h')$ for all $h' \in H$ where $|h'| = |h|+1$ and $h \sqsubseteq h'$ minimizes variance. Thus, having $V = V^\sigma$ everywhere minimizes variance. ■

Thus, if our knowledge of the dynamics is correct, then we know our estimator is unbiased (Theorem 1), and if our beliefs are correct, it minimizes variance (Theorem 3). However, what if our beliefs are not perfectly accurate? For instance, in poker, we can't perfectly predict the play of all the players. However, we might expect that in most situations the expected value under a belief and under the actual dynamics would be similar. We now show if we use a value function that is close to the true value function, then we get a random variable that is close to the minimum variance unbiased estimator.

Lemma 4 *For any $k \in \mathcal{K}$, $\sigma \in \Sigma_k$, and $V, V' : H \to \mathbf{R}$:*

$$\mathbf{E}_{h\in\sigma}\left[|\hat{u}_{V,k} - \hat{u}_{V',k}|\right] \le 2\sum_{h\in H} \Pr_\sigma[h]|V(h) - V'(h)|. \qquad (12)$$

Moreover, this closeness directly translates into a closeness in variance.

Theorem 5 *For any $k \in \mathcal{K}$, $\sigma \in \Sigma_k$, and $V, V' : H \to \mathbf{R}$, define $u_{\max} = \max_{h\in O}[\max(\hat{u}_{V,k}(h), \hat{u}_{V',k}(h))]$. It is the case that:*

$$\begin{aligned} &\mathbf{Var}_\sigma(\hat{u}_{V,k}) - \mathbf{Var}_\sigma(\hat{u}_{V',k}) \\ &\qquad \le 4u_{\max}\sum_{h\in H} \Pr_\sigma[h]|V(h) - V'(h)|. \end{aligned}$$

Thus, if on the histories we visit most we have a reasonably good estimate of the true value, and there is some trivial bound on how accurate we are on all possible histories, then we can be close to the optimal variance. In summary, if our knowledge of the dynamics is correct, then we know our estimator is unbiased, and if our beliefs are nearly correct, we'll have an estimator that has nearly the minimum variance.

Empirical Results

In the previous section we showed that knowledge of some portion of the system's dynamics as well as an accurate belief over the complete dynamics can lead to a low variance unbiased estimator. In this section we apply these results to the game of poker. Our knowledge consists of the rules of the game, i.e., we know the true distribution over the dealing and revealing of cards. Our belief must specify a guess of the expected outcome of the hand from any history. We've shown that one method for constructing such a function is to define a belief about the players' policies. The value function is then the expected value of the game if the players followed the chosen policies.

This is the approach of the Ignorant Value Assessment Tool (DIVAT) invented by the last author for assessing the value of poker decisions (Billings & Kan 2006). DIVAT makes use of an expert-defined policy for determining an appropriate amount players will wager in an arbitrary poker situation, called the DIVAT policy. The value function of this policy is then used in the advantage sum to make an unbiased estimator for poker called the DIVAT difference.

The DIVAT policy is based on a game-theoretic bet-for-value strategy. For example, if Player 1 holds a hand in the 70th percentile of strength and Player 2 holds a hand in 90th percentile, then the bet-for-value betting sequence would be bet, followed by a raise, followed by a call, indicating that each player should invest two bets on that betting round. The specific bet and raise thresholds are based on expected value equilibrium values, relative to a similarly defined game-theoretic equilibrium folding policy.

Implementation Details. To compute the estimator, one must compute the expected value of the DIVAT policy from various non-terminal histories. From a post-flop history, it requires a fraction of a second to compute the value, but from a pre-flop history, this computation can take over an hour. Therefore, we pre-compute and cache the value of all of the pre-flop histories, and then for later histories, we compute this value on the fly. On an AMD64 2.2Ghz machine, the analysis takes 0.418 seconds per hand on average.

Experiments. To evaluate our unbiased estimator in practice, we performed two experiments.[4] In both experiments we compare the DIVAT advantage sum estimator to the *money* estimator, based on averaging the player's per hand winnings. The first experiment was a self-play match with an experimental version of the advanced pseudo-optimal player (Billings *et al.* 2003). The particular program did not adapt to its opponent, so the expected winnings is zero. However, because of the stochasticity of poker, many hands are required to safely conclude this.

In our experiment evaluating seventy thousand hands, the money estimate has a standard deviation of 4.9 sb/h (small bets per hand). The DIVAT advantage sum estimator's standard deviation is 2.1 sb/h. In general, this means we would need 5.7 times the number of hands to have a money estimator with the same accuracy as the DIVAT estimator when evaluating this program in self-play. In Figure 1(a) we show both the estimated small bets per hand for the money and DIVAT estimators over the first two thousand hands of the experiment. The bars denote the 95% confidence interval given the sample standard deviation. The DIVAT advantage sum very quickly converges toward zero, while the money estimate is far less certain.

Our second experiment is a match between an expert poker player and the program that was used in the previous experiment. The expert used a fixed strategy he knew from prior experience would beat the program. For the ten thousand hands in the experiment, the money estimate had a standard deviation of 5.5 sb/h compared to DIVAT's 2.0 sb/h, resulting in 7.2 times fewer hands needed for similar accuracy. In Figure 1(b) we plot the same graph of estimators as in the self-play experiment. The money estimator requires approximately 800 hands before the break-even expected value is outside of its 95% confidence interval. It takes only 100 hands using the DIVAT advantage sum estimator to draw the same conclusion.

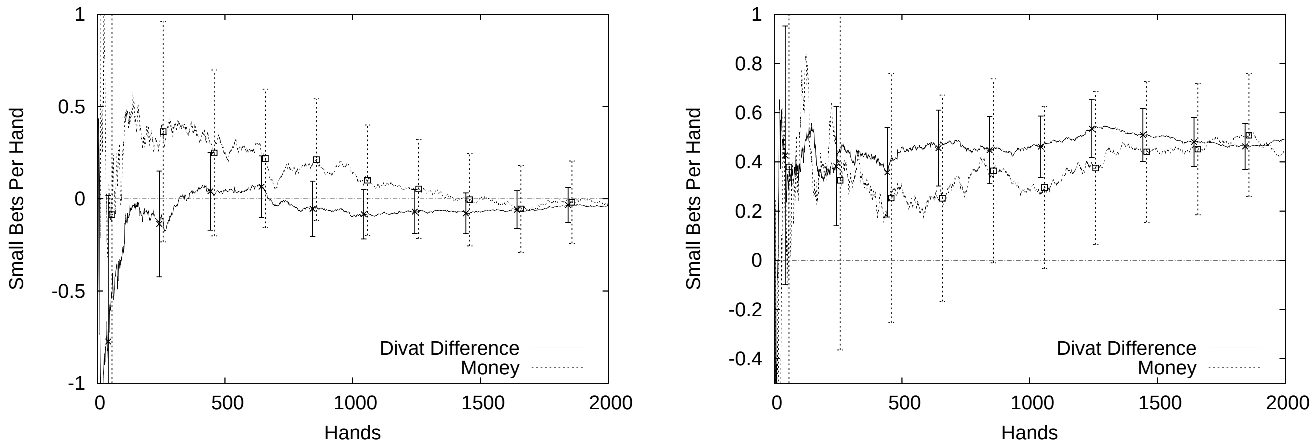
Hypothesis Testing. A common question in evaluation is simply, "On average, will Alice win money from Bob?" Or, "On average, will Alice win more from Bob than Charlie wins from Bob?" Given the results of an unbiased estimator this can be answered using hypothesis testing. Consider experiment two above, where we've seen just the first 500 hands and we want to ask, "On average, will the expert win money from the program?" A one-sided t-test using the DIVAT advantage sum estimator results in rejecting the null hypothesis that the human will break-even or lose to the program with a p-value less than 0.0001 (i.e., with a confidence level as high as 99.99%), which is extremely significant. Using the money estimate, we cannot reject the null hypothesis (p-value of 0.23) even with 90% confidence.

Similarly, suppose the observer does not know anything about the first program (A) in the first experiment, but knows that the second program (B) was the same one playing in experiment two. Now consider the question, "On average, will the expert win more from program B than program A will win from B?" Using the money estimator, after 500 hands, the null hypothesis that program A will win at least as much as the human cannot be rejected (p-value of 0.43). However, using the DIVAT estimator, the null hypothesis can be rejected with very high confidence (p-value of 0.002). In summary, the low variance of the DIVAT estimator results in more dramatic statistical conclusions.

Conclusion

We examined the problem of finding low variance unbiased estimators for evaluating agents in stochastic domains. We showed how to construct an unbiased estimator using advantage sums that exploits both partial knowledge about the system dynamics and a belief about the unknown dynamics. After giving a complete characterization of the space of unbiased estimators, we showed that if the belief is (nearly) accurate the estimator is (nearly) the minimum variance unbiased estimator. We then demonstrated the use of advantage sum estimators in the context of poker showing that the DI-

[4]Further experiments and poker analysis of DIVAT can be found in the technical report (Billings & Kan 2006).



(a) (b)

Figure 1: Unbiased estimators of performance over 2000 hand experiments. Vertical bars show the 95% confidence intervals for the estimators. (a) A static poker program in self-play. (b) An expert player against the static poker program.

VAT estimator has reduced variance and allows statistically significant conclusions to be drawn with much less data.

The advantage sum estimator has many applications, of which evaluating agents in stochastic multiagent scenarios is only one. Advantage sum estimators can also be used for policy evaluation or policy gradients in reinforcement learning (Kakade 2003). In this case, the domain knowledge actually consists of the agent's policy, and the unknown dynamics come from the environment's transition probabilities. Our results show that given a belief about the transition probabilities, a minimum variance unbiased estimator can be constructed. In addition, we can very naturally include additional knowledge about transition probabilities to improve the variance of this estimator. Unbiased estimators are also critical for online decision making algorithms. For example, Exp4 (Auer *et al.* 2002) is an algorithm for choosing among a set of suggested policies or experts. On each round, it selects a policy and observes a utility estimate. Its online guarantee does not require any assumptions of stationarity, but it does depend upon unbiased estimators of the chosen policy. More importantly, its practical performance depends critically on the variance of the estimators (Kocsis & Szepesvri 2005): the lower the variance, the stronger the performance.

Acknowledgments

We would like to thank the entire University of Alberta poker research group for their participation in preliminary discussions of this work. This research was supported by Alberta Ingenuity through the Alberta Ingenuity Centre for Machine Learning and iCore.

References

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. 2002. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing* 32(1):48–77.

Billings, D., and Kan, M. 2006. A tool for the direct assessment of poker decisions. Technical Report TR06-07, University of Alberta.

Billings, D.; Burch, N.; Davidson, A.; Holte, R.; Schaeffer, J.; Schauenberg, T.; and Szafron, D. 2003. Approximating game-theoretic optimal strategies for full-scale poker. In *Eighteenth International Joint Conference on Artificial Intelligence*.

Harsanyi, J. 1967. Games with incomplete information played by Bayesian players, parts I, II, and III. *Management Science* 14:159–182,320–334, and 486–502.

Kakade, S. 2003. *On the Sample Complexity of Reinforcement Learning*. Ph.D. Dissertation, Gatsby Computational Neuroscience Unit.

Kitano, H.; Kuniyoshi, Y.; Noda, I.; Asada, M.; Matsubara, H.; and Osawa, E. 1997. RoboCup: A challenge problem for AI. *AI Magazine* 18(1):73–85.

Kocsis, L., and Szepesvri, C. 2005. Reduced variance payoff estimation in adversarial bandit problems. In *ECML Workshop on Reinforcement Learning in Non-Stationary Environments*.

Sklansky, D. 1992. *The Theory of Poker*. Two Plus Two Publishing.

Stone, P., and Greenwald, A. 2005. The first international trading agent competition: Autonomous bidding agents. *Electronic Commerce Research* 5(2):229–265.

Wolfe, D. 2002. Distinguishing gamblers from investors at the blackjack table. In Schaeffer, J.; Müller, M.; and Björnsson, Y., eds., *Computers and Games 2002*, LNCS 2883, 1–10. Springer-Verlag.

Zinkevich, M.; Bowling, M.; Bard, N.; Kan, M.; and Billings, D. 2006. Optimal unbiased estimators for evaluating agent performance. Technical Report TR06-08, University of Alberta.