Strategy Evaluation in Extensive Games with Importance Sampling

Michael Bowling BOWLING@CS.UALBERTA.CA
Michael Johanson JOHANSON@CS.UALBERTA.CA
Neil Burch BURCH@CS.UALBERTA.CA
Duane Szafron DUANE@CS.UALBERTA.CA
Department of Computing Science, University of Alberta, Edmonton, Alberta, T6G 2E8 Canada

Abstract

Typically agent evaluation is done through Monte Carlo estimation. However, stochastic agent decisions and stochastic outcomes can make this approach inefficient, requiring many samples for an accurate estimate. We present a new technique that can be used to simultaneously evaluate many strategies while playing a single strategy in the context of an extensive game. This technique is based on importance sampling, but utilizes two new mechanisms for significantly reducing variance in the estimates. We demonstrate its effectiveness in the domain of poker, where stochasticity makes traditional evaluation problematic.

1. Introduction

Evaluating an agent's performance is a component of nearly all research on sequential decision making. Typically, the agent's expected payoff is estimated through Monte Carlo samples of the (often stochastic) agent acting in an (often stochastic) environment. The degree of stochasticity in the environment or agent behavior determines how many samples are needed for an accurate estimate of performance. For results in synthetic domains with artificial agents, one can simply continue drawing samples until the estimate is accurate enough. For non-synthetic environments, domains that involve human participants, or when evaluation is part of an on-line algorithm, accurate estimates with a small number of samples are critical. This paper describes a new technique for tackling this problem in the context of extensive games.

An extensive game is a formal model of a sequential interaction between multiple, independent agents with imperfect information. It is a powerful yet compact framework for describing many strategic interactions between decision-makers, artificial and human[1]. Poker, for example, is a domain modeled very naturally as an extensive game. It involves independent and self-interested agents making sequential decisions based on both public and private information in a stochastic environment. Poker also demonstrates the challenge of evaluating agent performance. In one typical variant of poker, approximately 30,000 hands (or samples of playing the game) are sometimes needed to distinguish between professional and amateur levels of play. Matches between computer and human opponents typically involve far fewer hands, yet still need to draw similar statistical conclusions.

In this work, we present a new technique for deriving low variance estimators of agent performance in extensive games. We employ importance sampling while exploiting the fact that the strategy of the agent being evaluated is typically known. However, we reduce the variance that importance sampling normally incurs by selectively adding synthetic data that is derived from but consistent with the sample data. As a result we derive low-variance unbiased estimators for agent performance given samples of the outcome of the game. We further show that we can efficiently evaluate one strategy while only observing samples from another. Finally, we examine the important case where we only get partial information of the game outcome (e.g., if a player folds in poker, their private cards are not revealed during the match and so the sequence of game states is not fully known). All of our estimators are then evaluated empirically in the domain of poker in both full and partial information scenarios.

This paper is organized as follows. In Section 2 we introduce the extensive game model, formalize our problem, and describe previous work on variance reduction in agent evaluation. In Section 3 we present a general procedure for deriving unbiased estimators and give four examples of

Appearing in *Proceedings of the 25^{th} International Conference on Machine Learning*, Helsinki, Finland, 2008.

[1] In this work we use the words "agent", "player", and "decision-maker" interchangeably and, unless explicitly stated, aren't concerned if they are humans or computers.

these estimators. We then briefly introduce the domain of poker in Section 4 and describe how these estimators can be applied to this domain. In Section 5 we show empirical results of our approach in poker. Finally, we conclude in Section 6 with some directions for future work.

2. Background

We begin by describing extensive games and then we formalize the agent evaluation problem.

2.1. Extensive Games

Definition 1 *(Osborne & Rubenstein, 1994, p. 200) a finite extensive game with imperfect information has the following components:*

- *A finite set N of* ***players****.*
- *A finite set H of sequences, the possible* ***histories*** *of actions, such that the empty sequence is in H and every prefix of a sequence in H is also in H. $Z \subseteq H$ are the* ***terminal histories*** *(those which are not a prefix of any other sequences). $A(h) = \{a : (h, a) \in H\}$ are the actions available after a non-terminal history $h \in H$,*
- *A* ***player function*** *P that assigns to each non-terminal history (each member of $H \backslash Z$) a member of $N \cup \{c\}$, where c represents chance. $P(h)$ is the player who takes an action after the history h. If $P(h) = c$, then chance determines the action taken after history h.*
- *A function f_c that associates with every history h for which $P(h) = c$ a probability measure $f_c(\cdot|h)$ on $A(h)$ ($f_c(a|h)$ is the probability that a occurs given h), where each such probability measure is independent of every other such measure.*
- *For each player $i \in N$ a partition $\mathbf{I}_i$ of $\{h \in H : P(h) = i\}$ with the property that $A(h) = A(h')$ whenever h and h' are in the same member of the partition. $\mathbf{I}_i$ is the* ***information partition*** *of player i; a set $I_i \in \mathbf{I}_i$ is an* ***information set*** *of player i.*
- *For each player $i \in N$ a utility function u_i from the terminal states Z to the reals $\mathbf{R}$. If $N = \{1, 2\}$ and $u_1 = -u_2$, it is a* ***zero-sum extensive game****.*

A **strategy of player i** σ_i in an extensive game is a function that assigns a distribution over $A(I_i)$ to each $I_i \in \mathbf{I}_i$. A **strategy profile** σ consists of a strategy for each player, $\sigma_1, \sigma_2, \ldots$, with σ_{-i} referring to all the strategies in σ except σ_i.

Let $\pi^\sigma(h)$ be the probability of history h occurring if players choose actions according to σ. We can decompose $\pi^\sigma = \Pi_{i \in N \cup \{c\}} \pi_i^\sigma(h)$ into each player's contribution to this probability. Hence, $\pi_i^\sigma(h)$ is the probability that if player i plays according to σ then for all histories h' that are a proper prefix of h with $P(h') = i$, player i takes the subsequent action in h. Let $\pi_{-i}^\sigma(h)$ be the product of all players' contribution (including chance) except player i. The overall value to player i of a strategy profile is then the expected payoff of the resulting terminal node, i.e., $u_i(\sigma) = \sum_{z \in Z} u_i(z)\pi^\sigma(z)$. For $Y \subseteq Z$, a subset of possible terminal histories, define $\pi^\sigma(Y) = \sum_{z \in Y} \pi^\sigma(z)$, to be the probability of reaching any outcome in the set Y given σ, with $\pi_i^\sigma(Y)$ and $\pi_{-i}^\sigma(Y)$ defined similarly.

2.2. The Problem

Given some function on terminal histories $V : Z \to \Re$ we want to estimate $E_{z|\sigma}[V(z)]$. In most cases V is simply u_i, and the goal is to evaluate a particular player's expected payoff. We explore three different settings for this problem. In all three settings, we assume that σ_i (our player's strategy) is known, while $\sigma_{j \neq i}$ (the other players' strategies) are not known.

- *On-policy full-information*. In the simplest case, we get samples $z_{1\ldots t} \in Z$ from the distribution π^σ.
- *Off-policy full-information*. In this case, we get samples $z_{1\ldots t} \in Z$ from the distribution $\pi^{\hat{\sigma}}$ where $\hat{\sigma}$ differs from σ only in player i's strategy: $\pi_{-i}^\sigma = \pi_{-i}^{\hat{\sigma}}$. In this case we want to evaluate one strategy for player i from samples of playing a different one.
- *Off-policy partial-information*. In the hardest case, we don't get full samples of outcomes z_t, but rather just player i's view of the outcomes. For example, in poker, if a player folds, their cards are not revealed to the other players and so certain chance actions are not known. Formally, in this case we get samples of $K(z_t) \in \mathbf{K}$, where K is a many-to-one mapping and z_t comes from the distribution $\pi^{\hat{\sigma}}$ as above. K intuitively must satisfy the following conditions: for $z, z' \in Z$, if $K(z) = K(z')$ then,
  - $V(z) = V(z')$, and
  - $\forall \sigma \quad \pi_i^\sigma(z) = \pi_i^\sigma(z')$.

2.3. Monte Carlo Estimation

The typical approach to estimating $E_{z|\sigma}[V(z)]$ is through simple Monte Carlo estimation. Given independent samples $z_1, \ldots, z_t$ from the distribution π^σ, simply estimate the expectation as the sample mean of outcome values.

$$\frac{1}{t}\sum_{i=1}^{t} V(z_i) \qquad (1)$$

As the estimator has zero bias, the mean squared error of the estimator is determined by its variance. If the variance of $V(z)$ given σ is large, the error in the estimate can be large and many samples are needed for accurate estimation.

Recently, we proposed a new technique for agent evaluation in extensive games (Zinkevich et al., 2006). We showed that value functions over non-terminal histories could be used to derive alternative unbiased estimators. If the chosen value function was close to the true expected value given the partial history and players' strategies, then the estimator would result in a reduction in variance. The approach essentially derives a real-valued function $\tilde{V}(z)$ that is used in place of V in the Monte Carlo estimator from Equation 1. The expectation of $\tilde{V}(z)$ matches the expectation of $V(z)$ for any choice of σ, and so the result is an unbiased estimator, but potentially with lower variance and thus lower mean-squared error. The specific application of this approach to poker, using an expert-defined value function, was named the DIVAT estimator and was shown to result in a dramatic reduction in variance. A simpler choice of value function, the expected value assuming the betting is "bet-call" for all remaining betting rounds, can even make a notable reduction. We refer to this conceptually and computationally simpler estimator as (Bet-Call) BC-DIVAT.

Both traditional Monte Carlo estimation and DIVAT are focused on the *on-policy* case, requiring outcomes sampled from the joint strategy that is being evaluated. Furthermore, DIVAT is restricted to *full-information*, where the exact outcome is known. Although limited in these regards, they also don't require any knowledge about any of the players' strategies.

3. General Approach

We now describe our new approach for deriving low-variance, unbiased estimators for agent evaluation. In this section we almost exclusively focus on the *off-policy full-information* case. Within this setting we observe a sampled outcome z from the distribution $\pi^{\hat{\sigma}}$, and the goal is to estimate $E_{z|\sigma}[V(z)]$. The outcomes are observed based on the strategy $\hat{\sigma}$ while we want to evaluate the expectation over σ, where they differ only in player i's strategy. This case subsumes the on-policy case, and we touch on the more difficult partial-information case at the end of this section. In order to handle this more challenging case, we require full knowledge of player i's strategies, both the strategy being observed $\hat{\sigma}_i$ and the one being evaluated σ_i.

At the core of our technique is the idea that synthetic histories derived from the sampled history can also be used in the estimator. For example, consider the unlikely case when σ is known entirely. Given an observed outcome $z \in Z$ (or even without an observed outcome) we can exactly compute the desired expectation by examining every outcome.

$$V_Z(z) \equiv \sum_{z' \in Z} V(z')\pi^{\sigma}(z') = E_{z|\sigma}[V(z)] \qquad (2)$$

Although impractical since we don't know σ, $V_Z(z)$ is an unbiased and zero variance estimator.

Instead of using every terminal history, we could restrict ourselves to a smaller set of terminal histories. Let $U(z' \in Z) \subseteq Z$ be a mapping of terminal histories to a set of terminal histories, where at least $z' \in U(z')$. We can construct an unbiased estimator that considers the history z' in the estimation whenever we observe a history from the set $U(z')$. Another way to consider things is to say that $U^{-1}(z)$ is the set of synthetic histories considered when we observe z. Specifically, we define the estimator $V_U(z)$ for the observed outcome z as,

$$V_U(z) \equiv \sum_{z' \in U^{-1}(z)} V(z') \frac{\pi^{\sigma}(z')}{\pi^{\hat{\sigma}}(U(z'))} \qquad (3)$$

The estimator considers the value of every outcome z' where the observed history z is in the set $U(z')$. Each outcome though is weighted in a fashion akin to importance sampling. The weight term for z' is proportional to the probability of that history given σ, and inversely proportional to the probability that z' is one of the considered synthetic histories when observing sampled outcomes from $\hat{\sigma}$. Note that $V_U(z)$ is not an estimate of $V(z)$, but rather has the same expectation.

At first glance, V_U may seem just as impractical as V_Z since σ is not known. However, with a careful choice of U we can insure that the weight term depends only on the known strategies σ_i and $\hat{\sigma}_i$. Before presenting example choices of U, we first prove that V_U is unbiased.

Theorem 1 *If $\pi_i^{\hat{\sigma}}(z)$ is non-zero for all outcomes $z \in Z$, then,*

$$E_{z|\hat{\sigma}}[V_U(z)] = E_{z|\sigma}[V(z)],$$

i.e., V_U is an unbiased estimator.

Proof: First, let us consider the denominator in the weight term of V_U. Since $z' \in U(z')$ and $\pi_i^{\hat{\sigma}}$ is always positive, the denominator can only be zero if $\pi_{-i}^{\hat{\sigma}}(z')$ is zero. If this were true, $\pi_{-i}^{\sigma}(z')$ must also be zero, and as a consequence so must the numerator. As a result the terminal history z' is never reached and so it is correct to simply exclude such histories from the estimator's summation.

Define $\mathbf{1}(x)$ to be the indicator function that takes on the value 1 if x is true and 0 if false.

$$E_{z|\hat{\sigma}}\left[V_U(z)\right]$$
$$= E_{z|\hat{\sigma}}\left[\sum_{z' \in U^{-1}(z)} V(z')\frac{\pi^{\sigma}(z')}{\pi^{\hat{\sigma}}(U(z'))}\right] \tag{4}$$
$$= E_{z|\hat{\sigma}}\left[\sum_{z'} \mathbf{1}(z \in U(z'))V(z')\frac{\pi^{\sigma}(z')}{\pi^{\hat{\sigma}}(U(z'))}\right] \tag{5}$$

$$= \sum_{z'} V(z')\frac{\pi^{\sigma}(z')}{\pi^{\hat{\sigma}}(U(z'))}E_{z|\hat{\sigma}}\left[\mathbf{1}(z \in U(z'))\right] \tag{6}$$
$$= \sum_{z'} V(z')\frac{\pi^{\sigma}(z')}{\pi^{\hat{\sigma}}(U(z'))}\pi^{\hat{\sigma}}(U(z')) \tag{7}$$
$$= \sum_{z'} V(z')\pi^{\sigma}(z') = E_{z|\sigma}\left[V(z)\right] \tag{8}$$

The derivation follows from the linearity of expectation, the definition of $\pi^{\hat{\sigma}}$, and the definition of expectation. ■

We now look at four specific choices of U for which the weight term can be computed while only knowing player i's portion of the joint strategy σ.

Example 1: Basic Importance Sampling. The simplest choice of U for which V_U can be computed is $U(z) = \{z\}$. In other words, the estimator considers just the sampled history. In this case the weight term is:

$$\frac{\pi^{\sigma}(z')}{\pi^{\hat{\sigma}}(U(z'))} = \frac{\pi^{\sigma}(z')}{\pi^{\hat{\sigma}}(z')} \tag{9}$$
$$= \frac{\pi_i^{\sigma}(z')\pi_{-i}^{\sigma}(z')}{\pi_i^{\hat{\sigma}}(z')\pi_{-i}^{\hat{\sigma}}(z')} \tag{10}$$
$$= \frac{\pi_i^{\sigma}(z')}{\pi_i^{\hat{\sigma}}(z')} \tag{11}$$

The weight term only depends on σ_i and $\hat{\sigma}_i$ and so is a known quantity. When $\hat{\sigma}_i = \sigma_i$ the weight term is 1 and the result is simple Monte Carlo estimation. When $\hat{\sigma}_i$ is different, the estimator is a straightforward application of importance sampling.

Example 2: Game Ending Actions. A more interesting example is to consider all histories that differ from the sample history by only a single action by player i and that action must be the last action in the history. For example, in poker, the history where the player being evaluated chooses to fold at an earlier point in the betting sequence is considered in this estimator. Formally, define $S_{-i}(z) \in H$ to be the shortest prefix of z where the remaining actions in z are all made by player i or chance. Let $U(z) = \{z' \in Z : S_{-i}(z) \text{ is a prefix of } z'\}$. The weight term becomes,

$$\frac{\pi^{\sigma}(z')}{\pi^{\hat{\sigma}}(U(z'))} = \frac{\pi^{\sigma}(z')}{\pi^{\hat{\sigma}}(S_{-i}(z'))} \tag{12}$$
$$= \frac{\pi_{-i}^{\sigma}(z')\pi_i^{\sigma}(z')}{\pi_{-i}^{\hat{\sigma}}(S_{-i}(z'))\pi_i^{\hat{\sigma}}(S_{-i}(z'))} \tag{13}$$
$$= \frac{\pi_{-i}^{\sigma}(S_{-i}(z'))\pi_i^{\sigma}(z')}{\pi_{-i}^{\hat{\sigma}}(S_{-i}(z'))\pi_i^{\hat{\sigma}}(S_{-i}(z'))} \tag{14}$$
$$= \frac{\pi_i^{\sigma}(z')}{\pi_i^{\hat{\sigma}}(S_{-i}(z'))} \tag{15}$$

As this only depends on the strategies of player i, we can compute this quantity and therefore the estimator.

Example 3: Private Information. We can also use all histories in the update that differ only in player i's private information. In other words, any history that the other players wouldn't be able to distinguish from the sampled history is considered. For example, in poker, any history where player i receiving different private cards is considered in the estimator since the opponents' strategy cannot depend directly on this strictly private information. Formally, let $U(z) = \left\{z' \in Z : \forall\sigma\ \pi_{-i}^{\sigma}(z') = \pi_{-i}^{\sigma}(z)\right\}$. The weight term then becomes,

$$\frac{\pi^{\sigma}(z')}{\pi^{\hat{\sigma}}(U(z'))} = \frac{\pi^{\sigma}(z')}{\sum_{z'' \in U(z')} \pi^{\hat{\sigma}}(z'')} \tag{16}$$
$$= \frac{\pi_{-i}^{\sigma}(z')\pi_i^{\sigma}(z')}{\sum_{z'' \in U(z')} \pi_{-i}^{\hat{\sigma}}(z'')\pi_i^{\hat{\sigma}}(z'')} \tag{17}$$
$$= \frac{\pi_{-i}^{\sigma}(z')\pi_i^{\sigma}(z')}{\sum_{z'' \in U(z')} \pi_{-i}^{\hat{\sigma}}(z')\pi_i^{\hat{\sigma}}(z'')} \tag{18}$$
$$= \frac{\pi_{-i}^{\sigma}(z')\pi_i^{\sigma}(z')}{\pi_{-i}^{\hat{\sigma}}(z')\sum_{z'' \in U(z')} \pi_i^{\hat{\sigma}}(z'')} \tag{19}$$
$$= \frac{\pi_i^{\sigma}(z')}{\pi_i^{\hat{\sigma}}(U(z'))} \tag{20}$$

As this only depends on the strategies of player i, we can again compute this quantity and therefore the estimator as well.

Example 4: Combined. The past two examples show that we can consider histories that differ in the player's private information or by the player making an alternative game ending action. We can also combine these two ideas and consider any history that differs by both an alternative game ending action and the player's private information. Define $Q(z) = \left\{h \in H : |h| = |S_{-i}(z)| \text{ and } \forall\sigma\pi_{-i}^{\sigma}(h) = \pi_{-i}^{\sigma}(S_{-i}(z))\right\}$, Let $U(z) = \{z' \in Z : \text{ a prefix of } z' \text{ is in } Q(z)\}$.

$$\frac{\pi^{\sigma}(z')}{\pi^{\hat{\sigma}}(U(z'))} = \frac{\pi^{\sigma}(z')}{\pi^{\hat{\sigma}}(Q(z'))} \tag{21}$$

$$= \frac{\pi^{\sigma}_{-i}(z')\pi^{\sigma}_{i}(z')}{\sum_{h \in Q(z')} \pi^{\hat{\sigma}}_{-i}(h)\pi^{\hat{\sigma}}_{i}(h)} \tag{22}$$

$$= \frac{\pi^{\sigma}_{-i}(z')\pi^{\sigma}_{i}(z')}{\sum_{h \in Q(z')} \pi^{\hat{\sigma}}_{-i}(S_{-i}(z))\pi^{\hat{\sigma}}_{i}(h)} \tag{23}$$

$$= \frac{\pi^{\sigma}_{-i}(S_{-i}(z'))\pi^{\sigma}_{i}(z')}{\pi^{\hat{\sigma}}_{-i}(S_{-i}(z')) \sum_{h \in Q(z')} \pi^{\hat{\sigma}}_{i}(h)} \tag{24}$$

$$= \frac{\pi^{\sigma}_{i}(z')}{\pi^{\hat{\sigma}}_{i}(Q(z'))} \tag{25}$$

Once again this quantity only depends on the strategies of player i and so we can compute this estimator as well.

We have presented four different estimators that try to extract additional information from a single observed game outcome. We can actually combine any of these estimators with other unbiased approaches for reducing variance. This can be done by replacing the V function in the above estimators with any unbiased estimate of V. In particular, these estimators can be combined with our previous DIVAT approach by choosing V to be the DIVAT (or BC-DIVAT) estimator instead of u_i.

3.1. Partial Information

The estimators above are provably unbiased for both the-policy and off-policy full-information case. We now briefly discuss the off-policy partial-information case. In this case we don't directly observe the actual terminal history z_t but only a many-to-one mapping $K(z_t)$ of the history. One simple adaptation of our estimators to this case is to use the history z' in the estimator whenever it is possible that the unknown terminal history could be in $U(z')$, while keeping the weight term unchanged. Although we lose the unbiased guarantee with these estimators, it is possible that the reduction in variance is more substantial than the error caused by the bias. We investigate empirically the magnitude of the bias and the resulting mean-squared error of such estimators in the domain of poker in Section 5.

4. Application to Poker

To analyze the effectiveness of these estimators, we will use the popular game of Texas Hold'em poker, as played in the AAAI Computer Poker Competition (Zinkevich & Littman, 2006). The game is two-player and zero-sum. Private cards are dealt to the players, and over four rounds, public cards are revealed. During each round, the players place bets that the combination of their public and private cards will be the strongest at the end of the game. The game has just under 10^{18} game states, and has the properties of imperfect information, stochastic outcomes, and observations of the game outcome during a match exhibit partial information.

Each of the situations described in Section 2, on-policy and off-policy as well as full-information and partial information, have relevance in the domain of poker. In particular, the *on-policy full-information* case is the situation where one is trying to evaluate a strategy from full-information descriptions of the hands, as might be available after a match is complete. For example, this could be used to more accurately determine the winner of a competition involving a small number of hands (which is always the case when humans are involved). In this situation it is critical, that the estimator is unbiased, i.e., it is an accurate reflection of the expected winnings and therefore does not incorrectly favor any playing style.

The *off-policy full-information* case is useful for examining past games against an opponent to determine which of many alternative strategies one might want to use against them in the future. The introduction of bias (depending on the strategy used when playing the past hands) is not problematic, as the goal in this case is an estimate with as little error as possible. Hence the introduction of bias is acceptable in exchange for significant decreases in variance.

Finally, the *off-policy partial-information* case corresponds to evaluating alternative strategies during an actual match. In this case, we want to evaluate a set of strategies, which aren't being played, to try and identify an effective choice for the current opponent. The player could then choose a strategy whose performance is estimated to be strong even for hands it wasn't playing.

The estimators from the previous section all have natural applications to the game of poker:

- **Basic Importance Sampling**. This is a straightforward application of importance sampling. The value of the observed outcome of the hand is weighted by the ratio of the probability that the strategy being evaluated (σ_i) takes the same sequence of actions to the probability that the playing strategy ($\hat{\sigma}_i$) takes the sequence of actions.

- **Game ending actions**. By selecting the *fold* betting action, a player surrenders the game in order to avoid matching an opponent's bet. Therefore, the game ending actions estimator can consider all histories in which the player could have folded during the observed history.[2] We call this the **Early Folds** (EF) estimator. The estimator sums over all possible prefixes

[2] In the full-information setting we can also consider situations where the player could have *called* on the final round of betting to end the hand.

of the betting sequence where the player could have chosen to fold. In the summation it weights the value of surrendering the pot at that point by the ratio of the probability of the observed betting up to that point and then folding given the player's cards (and σ_i) to the probability of the observed betting up to that point given the player's cards (and $\hat{\sigma}_i$).

- **Private information**. In Texas Hold'em, a player's private information is simply the two private cards they are dealt. Therefore, the private information estimator can consider all histories with the same betting sequence in which the player holds different private cards. We call this the **All Cards** (AC) estimator. The estimator sums over all possible two-card combinations (excepting those involving exposed board or opponent cards). In the summation it weights the value of the observed betting with the imagined cards by the ratio of the probability of the observed betting given those cards (and σ_i) to the probability of the observed betting (given $\hat{\sigma}_i$) summed over all cards.

5. Results

Over the past few years we have created a number of strong Texas Hold'em poker agents that have competed in the past two AAAI Computer Poker Competitions. To evaluate our new estimators, we consider games played between three of these poker agents: S2298 (Zinkevich et al., 2007), PsOpti4 (Billings et al., 2003), and CFR8 (Zinkevich et al., 2008). In addition, we also consider Orange, a competitor in the First Man-Machine Poker Championship.

To evaluate these estimators, we examined records of games played between each of three candidate strategies (S2298, CFR8, Orange) against the opponent PsOpti4. Each of these three records contains one million hands of poker, and can be viewed as full information (both players' private cards are always shown) or as partial information (when the opponent folds, their private cards are not revealed). We begin with the full-information experiments.

5.1. Full Information

We used the estimators described previously to find the value of each of the three candidate strategies, using full-information records of games played from just one of the candidate strategies. The strategy that actually played the hands in the record of games is called the on-policy strategy and the others are the off-policy strategies. The results of one these experiments is presented in Table 1. In this experiment, we examined one million full-information hands of S2298 playing against PsOpti4. S2298 (the on-policy strategy) and CFR8 and Orange (the off-policy strategies) are evaluated by our importance sampling estimators, as well as DIVAT, BC-DIVAT, and a few combination estimators. We present the empirical bias and standard deviation of the estimators in the first two columns. The third column, "RMSE", is the root-mean-squared error of the estimator if it were used as the method of evaluation for a 1000 hand match (a typical match length). All of the numbers are reported in millibets per hand played. A millibet is one thousandth of a small-bet, the fixed magnitude of bets used in the first two rounds of betting. To provide some intuition for these numbers, a player that always folds will lose 750 millibets per hand, and strong players aim to achieve an expected win rate over 50 millibets per hand.

Table 1. Full Information Case. Empirical bias, standard deviation, and root mean-squared-error over a 1000 hand match for various estimators. 1 million hands of poker between S2298 and PsOpti4 were observed. A bias of 0* indicates a provably unbiased estimator.

	Bias	StdDev	RMSE
S2298			
Basic	0*	5103	161
DIVAT	0*	1935	61
BC-DIVAT	0*	2891	91
Early Folds	0*	5126	162
All Cards	0*	4213	133
AC+BC-DIVAT	0*	2146	68
AC+EF+BC-DIVAT	0*	1778	56
CFR8			
Basic	200 ± 122	62543	1988
DIVAT	62 ± 104	53033	1678
BC-DIVAT	84 ± 45	22303	710
Early Folds	123 ± 120	61481	1948
All Cards	12 ± 16	8518	270
AC+BC-DIVAT	35 ± 13	3254	109
AC+EF+BC-DIVAT	2 ± 12	2514	80
Orange			
Basic	159 ± 40	20559	669
DIVAT	3 ± 25	11350	359
BC-DIVAT	103 ± 28	12862	420
Early Folds	82 ± 35	17923	572
All Cards	7 ± 16	8591	272
AC+BC-DIVAT	8 ± 13	3154	100
AC+EF+BC-DIVAT	6 ± 12	2421	77

In the on-policy case, where we are evaluating S2298, all of the estimators are provably unbiased, and so they only differ in variance. Note that the Basic estimator, in this case, is just the Monte-Carlo estimator over the actual money lost or won. The Early Folds estimator provides no variance reduction over the Monte-Carlo estimate, while the All Cards estimator provides only a slight reduction. However, this is not nearly as dramatic as the reduction provided by the DIVAT estimator. The importance sampling estimators, however, can be combined with the DIVAT estimator as described in Section . The combination of BC-DIVAT with All Cards ("AC+BC-DIVAT") results in lower variance than either of the estimators separately.[3] The addition of Early Folds ("AC+EF+BC-DIVAT") produces an even further reduction in variance, showing the best-performance of all the estimators, even though Early Folds on its own had little effect.

In the off-policy case, where we are evaluating CFR8 or Orange, we report the empirical bias (along with a 95% confidence bound) in addition to the variance. As DIVAT and BC-DIVAT were not designed for off-policy evaluation, we report numbers by combining them with the Basic estimator (i.e., using traditional importance sampling). Note that bias is possible in this case because our on-policy strategy (S2298) does not satisfy the assumption in Theorem 1, as there are some outcomes the strategy never plays. Basic importance sampling in this setting not only shows statistically significant bias, but also exhibits impractically large variance. DIVAT and BC-DIVAT, which caused considerable variance reduction on-policy, also should considerable variance reduction off-policy, but not enough to offset the extra variance from basic importance sampling. The All Cards estimator, on the other hand, shows dramatically lower variance with very little bias (in fact, the empirical bias is statistically insignificant). Combining the All Cards estimator with BC-DIVAT and Early Folds further reduces the variance, giving off-policy estimators that are almost as accurate as our best on-policy estimators.

The trends noted above continue in the other experiments, when CFR8 and Orange are being observed. For space considerations, we don't present the individual tables, but instead summarize these experiments in Table 2. The table shows the minimum and maximum empirically observed bias, standard deviation, and the root-mean-squared error of the estimator for a 1000 hand match. The strategies being evaluated are separated into the on-policy case, when the record involves data from that strategy, and the off-policy case, when it doesn't.

5.2. Partial Information

The same experiments were repeated for the case of partial information. The results of the experiment involving S2298 playing against PsOpti4 and evaluating our three candidate strategies under partial information is shown in Table 3. For DIVAT and BC-DIVAT, which require full information of the game outcome, we used a partial information variant where the full-information estimator was used when the game outcome was known (i.e., no player folded) and winnings was used when it was not. This variant can result in a biased estimator, as can be seen in the table of results. The All Cards estimator, although also without any guarantee of being unbiased, actually fares much better in practice, not displaying a statistically significant bias in either the off-policy or on-policy experiments. However, even though the DIVAT estimators are biased their low variance makes them preferred in terms of RMSE in the on-policy setting. In the off-policy setting, the variance caused by Basic importance sampling (as used with DIVAT and BC-DIVAT) makes the All Cards estimator the only practical choice. As in the full-information case we can combine the All Cards and BC-DIVAT for further variance reduction. The resulting estimator has lower RMSE than either All Cards or BC-DIVAT alone both in the on-policy and off-policy cases. The summary of the results of the other experiments, showing similar trends, are shown in Table 4.

Table 3. Partial-Information Case. Empirical bias, standard deviation, and root mean-squared-error over a 1000 hand match for various estimators. 1 million hands of poker between S2298 and PsOpti4 with partial information were observed. A bias of 0* indicates a provably unbiased estimator.

	Bias	StdDev	RMSE
S2298			
Basic	0*	5104	161
DIVAT	81±9	2762	119
BC-DIVAT	95±9	2759	129
Early Folds	47±1	5065	167
All Cards	5±13	4218	133
AC+BC-DIVAT	96±12	2650	127
CFR8			
Basic	202±80	40903	1309
DIVAT	175±47	23376	760
BC-DIVAT	183±47	23402	762
Early Folds	181±78	39877	1274
All Cards	13±19	7904	250
AC+BC-DIVAT	101±16	4014	162
Orange			
Basic	204±45	23314	765
DIVAT	218±22	10029	385
BC-DIVAT	244±21	10045	401
Early Folds	218±43	22379	741
All Cards	3±19	8092	256
AC+BC-DIVAT	203±16	3880	237

6. Conclusion

We introduced a new method for estimating agent performance in extensive games based on importance sampling. The technique exploits the fact that the agent's strategy is typically known to derive several low variance estima-

[3]The importance sampling estimators were combined with BC-DIVAT instead of DIVAT because the original DIVAT estimator is computationally burdensome, particularly when many evaluations are needed for every observation as is the case with the All Cards estimator.

Table 2. Summary of the Full-Information Case. Summary of empirical bias, standard deviation, and root-mean-squared error over a 1000 hand match for various estimators. The minimum and maximum encountered values for all combinations of observed and evaluated strategies is presented. A bias of 0* indicates a provably unbiased estimator.

	Bias			StdDev			RMSE		
	Min	–	Max	Min	–	Max	Min	–	Max
On Policy									
Basic	0*	–	0*	5102	–	5385	161	–	170
DIVAT	0*	–	0*	1935	–	2011	61	–	64
BC-DIVAT	0*	–	0*	2891	–	2930	91	–	92
AC+GE+BC-DIVAT	0*	–	0*	1701	–	1778	54	–	56
Off Policy									
Basic	49	–	200	20559	–	244469	669	–	7732
DIVAT	2	–	62	11350	–	138834	358	–	4390
BC-DIVAT	10	–	103	12862	–	173715	419	–	5493
AC+GE+BC-DIVAT	2	–	9	1816	–	2857	58	–	90

Table 4. Summary of the Partial-Information Case. Summary of empirical bias, standard deviation, and root-mean-squared error over a 1000 hand match for various estimators. The minimum and maximum encountered values for all combinations of observed and evaluated strategies is presented. A bias of 0* indicates a provably unbiased estimator.

	Bias			StdDev			RMSE		
	Min	–	Max	Min	–	Max	Min	–	Max
On Policy									
Basic	0*	–	0*	5104	–	5391	161	–	170
DIVAT	56	–	144	2762	–	2876	105	–	170
BC-DIVAT	78	–	199	2759	–	2859	118	–	219
AC+BC-DIVAT	78	–	206	2656	–	2766	115	–	224
Off Policy									
Basic	17	–	433	23314	–	238874	753	–	7566
DIVAT	103	–	282	10029	–	88791	384	–	2822
BC-DIVAT	35	–	243	10045	–	99287	400	–	3139
AC+BC-DIVAT	63	–	230	3055	–	6785	143	–	258

tors that can simultaneously evaluate many strategies while playing a single strategy. We prove that these estimators are unbiased in both the on-policy and off-policy case. We empirically evaluate the techniques in the domain of poker, showing significant improvements in terms of lower variance and lower bias. We show that the estimators can also be used even in the challenging problem of estimation with partial information observations.

Acknowledgments

We would like to thank Martin Zinkevich and Morgan Kan along with the entire University of Alberta Computer Poker Research Group for their valuable insights. This research was supported by NSERC and iCore.

References

Billings, D., Burch, N., Davidson, A., Holte, R., Schaeffer, J., Schauenberg, T., & Szafron, D. (2003). Approximating game-theoretic optimal strategies for full-scale poker. *International Joint Conference on Artificial Intelligence* (pp. 661–668).

Osborne, M., & Rubenstein, A. (1994). *A course in game theory*. Cambridge, Massachusetts: The MIT Press.

Zinkevich, M., Bowling, M., Bard, N., Kan, M., & Billings, D. (2006). Optimal unbiased estimators for evaluating agent performance. *American Association of Artificial Intelligence National Conference, AAAI'06* (pp. 573–578).

Zinkevich, M., Bowling, M., & Burch, N. (2007). A new algorithm for generating equilibria in massive zero-sum games. *Proceedings of the Twenty-Second Conference on Artificial Intelligence* (pp. 788–793).

Zinkevich, M., Johanson, M., Bowling, M., & Piccione, C. (2008). Regret minimization in games with incomplete information. *Advances in Neural Information Processing Systems 20*. To appear (8 pages).

Zinkevich, M., & Littman, M. (2006). The AAAI computer poker competition. *Journal of the International Computer Games Association*, *29*. News item.