The Baseline Approach to Agent Evaluation

Josh Davidson and **Christopher Archibald** and **Michael Bowling**
{joshuad, archibal, bowling}@ualberta.ca
Department of Computing Science
University of Alberta
Edmonton, AB, Canada T6G 2E8

Abstract

An important aspect of agent evaluation in stochastic games, especially poker, is the need to reduce the outcome variance in order to get accurate and significant results. The current method used in the Annual Computer Poker Competition's analysis is that of duplicate poker, an approach that leverages the ability to deal sets of cards to agents in order to reduce variance. This work explores a different approach to variance reduction by using a control variate based approach known as baseline. The baseline approach involves using an agent's outcome in self play to create an unbiased estimator for use in agent evaluation and has been shown to work well in both poker and trading agent competition domains. Baseline does not require that the agents are able to be dealt sets of cards, making it a more robust technique than duplicate. This approach is compared to the current duplicate method, as well as other variations of duplicate poker on the results of the 2011 two player no-limit and three player limit Texas Hold'em ACPC tournaments.

Introduction

Effective methods of agent evaluation is a critical component of the Annual Computer Poker Competition (ACPC). Due to the high variance inherent in the variants of Texas Hold'em that are played in the competition, techniques that reduce outcome variance are an important piece of the evaluation mechanism. This is increasingly true as the competitors' agents become better and better and more samples are needed to separate out the top agents from one another. Figure 1 shows just how bad this can be in a simple case. The graph demonstrates that it can take thousands of hands to separate an agent that always calls against an agent that always raises, when using only raw utility as the estimate of skill. As well, the actual expected value of these two agents is zero, and the graph shows always call winning with significance after 3000 hands. The solution to simply play more matches is not a viable option, especially with the move to pay-as-you-go services such as Amazon's EC2 or other cloud based services to run the competitions.

Copyright © 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

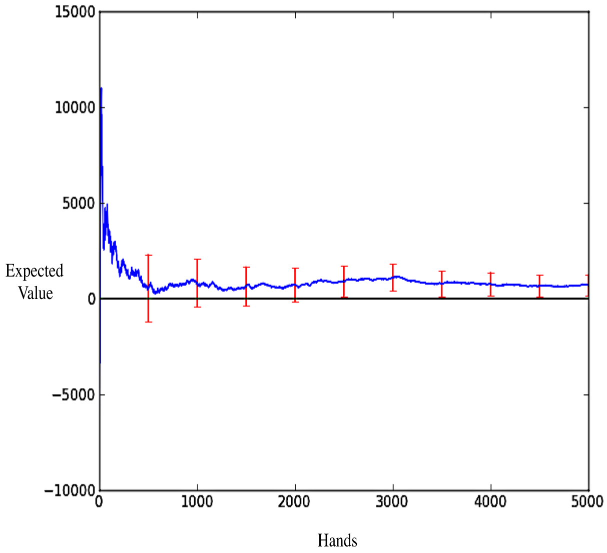

Figure 1: Expected Value of Always Call vs Always Raise

Current Approaches to Variance Reduction

Traditionally the ACPC tournaments have approached the variance problem by using duplicate poker. The power of duplicate comes from the ability to average the payouts over pairs, or sets, of hands where the cards remain the same and the positions of the agents are permuted. This approach is quite powerful when the outcomes of the permutations are strongly negatively correlated and has proven an effective method for use in the competition. Other approaches that have been studied such as DIVAT (Billings and Kan 2006), MIVAT (White and Bowling 2009) have been shown to provide greater variance reduction than duplicate. These approaches are more generally known as a advantage sum estimators (Zinkevich, Bowling, and Bard 2006) and are essentially trying to remove the effects of the positional and luck terms of the outcome, leaving only the skill to be evaluated. One problem with advantage sum estimators is they often require a value function, either hand coded or machine learned in some way. As well, they often require more computational resources than methods such as duplicate and hence are less frequently used. The baseline approach focuses on being a simple alternative to duplicate by employing a control variate based approach to variance reduction.

The Baseline Approach

The baseline approach is a method of reducing the outcome variance in zero sum domains. Baseline is a way to adjust the payouts observed by the agent you are evaluating. The adjusted payouts can then be used as an unbiased estimate of the agents skill. The baseline approach achieves this by effectively creating a control variate.

Control Variates

Control variates are a method of variance reduction that rely on the correlation between different random variables (Glasserman 1962). Classical approaches to control variates involve computing a statistic that correlates well with the outcome you are estimating. Given control variate Y and a random variable X, let Z be defined as

$$Z := X + c(Y - \mathbb{E}[Y]) \tag{1}$$

The sample mean of Z is then equal to:

$$\begin{aligned} \bar{Z} &= \bar{X} + c(\bar{Y} - \mathbb{E}[Y]) \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i + c(Y_i - \mathbb{E}[Y])) \end{aligned} \tag{2}$$

Proof that $\mathbb{E}[Z]$ is an unbiased estimator of $\mathbb{E}[X]$ for any $c \in \mathbb{R}$ is given by the following:

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}[X + c(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[X] + \mathbb{E}[c(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[X] + c(\mathbb{E}[Y] - \mathbb{E}[Y]]) \\ &= \mathbb{E}[X] \end{aligned} \tag{3}$$

The variance of Z can then be computed as

$$\begin{aligned} \mathbb{V}\text{ar}[Z] &= \mathbb{V}\text{ar}[X + c(Y - \mathbb{E}[Y])] \\ &= \mathbb{V}\text{ar}[X] + c^2\mathbb{V}\text{ar}[Y] + 2c\mathbb{C}\text{ov}[X, Y] \end{aligned} \tag{4}$$

The optimal coefficient c^*, which is the one that minimizes the variance, is then defined as

$$c^* = -\frac{\mathbb{C}\text{ov}[X, Y]}{\mathbb{V}\text{ar}[X]} \tag{5}$$

Which then gives the minimal variance equation for $\mathbb{V}\text{ar}[Z]$

$$\begin{aligned} \mathbb{V}\text{ar}[Z] &= \mathbb{V}\text{ar}[X] - \frac{\mathbb{C}\text{ov}[X, Y]^2}{\mathbb{V}\text{ar}[Y]} \\ &= (1 - \rho^2_{X,Y})\mathbb{V}\text{ar}[X] \end{aligned} \tag{6}$$

where $\rho_{X,Y}$ is the correlation between X and Y.

Baseline Control Variates

The typical difficulty with control variates is that the expected value of Y needs to be known or easily computable. The baseline approach eliminates the need to compute this expected value, thus eliminating the one of the encumbrances of using control variates (Davidson, Archibald, and Bowling 2013). Baseline control variates are created by first running an agent in self play on the same random events observed by the agent being evaluated. For poker, this simply is the deal of the cards, thus baseline does require full information of the environment. A baseline score is then computed for each observation, which is the Y term in the control variate equations. The trick to baseline is that the expected value of Y is zero. This is due to the fact that for a zero-sum game with rotating position, an agent in self play has an expected value of zero. Now we can simply subtract the baseline agent's score from those of the agent we are evaluating. These adjusted payouts can now be used to compute a lower variance, unbiased estimate of the agent's performance.

One thing to note is that is that any agent can be used to generate the payout adjustments. The agent in question does need to be a particularly strong agent, as the magnitude of the variance reduction is hinged on the correlation between the baseline scores and the observed payouts. Not only is baseline a simple alternative to duplicate for agent versus agent competitions, but baseline can be used in situations where duplicate is not possible, such as matches against human opponents.

Evaluation Methodology

For this work, the results from the 2011 Annual Computer Poker Competition's three player limit Texas Hold'em were studied. The baseline method was applied to these competitions using an agent that did not compete in that years competition. The baseline utilities were run on the same seeds as the competition, using a 50 sample average as the baseline score for each hand. In addition to the baseline analysis, an extensive analysis of the duplicate method was employed on the three player limit results. This analysis involved doing duplicate in the traditional ACPC approach of using all permutations of deals as well as separating out the rotational permutations from those where the seating was flipped.

Results

Tables 1 and 2 show the results of the analysis applied to the 2011 three player limit competition. The MEAN column of the table refers to the player's average utility with the 95% CI column representing the 95% confidence interval on this mean. All other columns of the table show the reduction in the variance the various estimators achieve, displayed as a percentage. The baseline estimator is labeled BASE, and the duplicate estimators corresponding to rotation, full and flip duplicate are lab led D-ROT, D-FULL and D-FLIP respectively. The larger this percentage, the greater the reduction of variance. Negative number in these tables represent an increase in the variance.

Table 1: 2011 ACPC Three Player Limit Texas Hold'em

Player	MEAN	95% CI	D-ROT (%)	D-FLIP (%)	BASE (%)	D-FULL (%)
dcubot3plr	62.93	4.38	**27.76**	-25.71	26.68	21.69
Bnold3	-103.78	4.05	**27.81**	-8.56	25.99	24.58
OwnBot	-15.42	5.22	**27.7**	12.77	23.54	19.99
player_zeta	-512.35	6.1	**17.56**	-19.59	16.17	1.55
Entropy	-23.95	6.01	**21.64**	7.16	21.12	11.25
LittleRock	105.86	3.79	**31.64**	9.75	29.56	28.2
Sartre3p	232.46	3.95	**33.7**	17.43	31.36	30.82
Hyper-iro	208.25	4.62	**31.33**	1.34	31.11	28.06
Hyper-tbr	170.15	5.25	**29.29**	-3.92	28.08	26.28

Table 2: 2011 ACPC Three Player Limit Texas Hold'em dcubot3plr results

Player	MEAN	95% CI	BASE (%)	D-ROT (%)	D-FULL (%)	D-FLIP (%)
dcubot3plr.player_zeta.Entropy	-5.64	7.73	**14.85**	12.86	3.69	-19.38
dcubot3plr.Bnold3.Hyper-iro	1.34	5.58	37.86	**41.17**	36.75	-30.89
dcubot3plr.player_zeta.Sartre3p	25.18	6.56	**22.38**	20.92	14.61	-22.33
dcubot3plr.player_zeta.OwnBot	9.13	7.55	**15.14**	14.56	1.77	-21.61
dcubot3plr.Bnold3.Entropy	12.56	7.03	27.54	**29.9**	23.48	-26.28
dcubot3plr.Bnold3.OwnBot	11.02	6.02	27.19	**29.44**	29.38	-24.22
dcubot3plr.Hyper-iro.Sartre3p	-6.61	5.42	41.45	**44.27**	38.39	-32.96
dcubot3plr.Entropy.Sartre3p	7.96	6.87	30.59	**32.5**	26.32	-27.49
dcubot3plr.player_zeta.Hyper-iro	22.83	6.82	**23.54**	22.24	14.87	-23.31
dcubot3plr.player_zeta.LittleRock	26.7	6.6	**22.61**	21.02	14.75	-22.37
dcubot3plr.Bnold3.LittleRock	5.46	5.39	36.38	**42.57**	37.1	-32.02
dcubot3plr.Hyper-iro.Entropy	6.25	6.98	30.81	**32.25**	26.07	-27.26
dcubot3plr.Bnold3.Sartre3p	1.93	5.46	38.32	**44.12**	39.58	-32.08
dcubot3plr.Sartre3p.OwnBot	4.21	5.83	29.38	**32.59**	31.52	-26.55
dcubot3plr.LittleRock.Entropy	8.93	6.86	30.04	**33.09**	28.76	-26.96
dcubot3plr.Hyper-iro.OwnBot	5.71	6.0	29.59	**30.8**	27.48	-26.62
dcubot3plr.Hyper-iro.LittleRock	-6.42	5.4	40.32	**43.25**	36.63	-32.63
dcubot3plr.player_zeta.Bnold3	25.97	6.93	**21.07**	20.05	13.19	-21.53
dcubot3plr.LittleRock.Sartre3p	-4.79	5.29	40.44	**47.08**	40.21	-34.02

Shown in Table 1 is a summary of the 2011 three player limit competitions. The summary shows the average across all opponents for the player being evaluated. Two important observations can be made from this table. The first of which is that the D-ROT column provides the best estimator in the three player competition for all of the summaries. This suggest that doing rotation duplicate should replace the current full duplicate approach in three player limit. This is not surprising however given that the positional portion of the averaging appears to provide extra variance, shown by the D-FLIP columns containing negative numbers. The second observation is that baseline (BASE) estimator comes in a fairly close second in all of the match summaries and is better than the full duplicate approach. One advantage the baseline approach has in this sense is that the baseline estimates can be run in advance before the competition. This simplifies the analysis a slight bit in that the seat rotations need not be executed, nor the seat permutations, with little loss in variance reduction. The agents can just be sat at a table and given random cards and deals according to the pre-computed baseline seeds.

Table 2 shows all of the matches involving dcubot3plr. These results show that on the non-aggregate data, the BASE column can out-perform D-ROT on individual matches. Here, the D-FULL column rarely out performs BASE and never outperforms D-ROT. As well, in the cases where BASE loses out to D-ROT or D-FULL, there are very few instances that the margin is large. Overall, baseline appears to perform nearly as well as D-ROT, both of which providing similar magnitudes of variance reduction.

Conclusions

The baseline approach provides a way of creating unbiased estimators for use in not only poker, but any zero-sum domain. The baseline estimators work well in both no-limit and three player poker games, providing estimates with reduced variance in both domains. The three player limit results suggest that a different form of variance reduction should be explored and used in the competition, either that of baseline or a different duplicate method.

Future Work

There are some obvious extensions to the baseline approach that were not explored in this work. Eliminating the need to sample the baseline strategy when computing the estimate is worth exploring, especially for domains where the control agents might act slowly. This would eliminate any noise in the estimate, which would presumably give better baseline scores for use in estimation. Another possible use of baseline is as a way to help aid learning approaches by providing lower variance utilities for use in online adaptation. As well, baseline could be used as an approach not only in poker, but other domains. Exploration of the usefulness of baseline in other high variance zero-sum domains would be an interesting and valuable study. Lastly, the baseline approach could also be used in systems to provide lower variance skill estimates in human versus agent and human versus human competitions.

Acknowledgments

We would like to thank the University of Alberta Poker Research Group for the use of their poker agents and thoughtful discussion. This research was supported by NSERC and Alberta Innovates through the Alberta Innovates Centre for Machine Learning.

References

Billings, D., and Kan, M. 2006. A tool for the direct assessment of poker decisions. *The International Computer Games Association Journal,* 29(3):119–142.

Davidson, J.; Archibald, C.; and Bowling, M. 2013. Baseline: Practical Control Variates for Agent Evaluation in Zero-Sum Domains. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems, AAMAS'13*, 1005–1012.

Glasserman, P. 1962. *Monte Carlo Methods in Financial Engineering*. Springer.

White, M., and Bowling, M. 2009. Learning a value analysis tool for agent evaluation. *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI)* 1976 – 1981.

Zinkevich, M.; Bowling, M.; and Bard, N. 2006. Optimal unbiased estimators for evaluating agent performance. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI)*, 573–578.