Strategy Evaluation in Extensive Games with Importance Sampling

Michael Bowling, Michael Johanson, Neil Burch, Duane Szafron

July 8, 2008

UNIVERSITY OF ALBERTA

University of Alberta
Computer Poker Research Group

Making IT happen Computing Science

Second Man-Machine Poker Championship

- Just arrived from the Second Man-Machine Poker Championship in Las Vegas
- Our program, Polaris, played six 500 hand duplicate matches against six poker pros over 4 days
- Final score: 3 wins, 2 losses, 1 tie! AI Wins!
- This research played a critical role in our success

The Problem

- Several candidate strategies to choose from
- Only have samples of one strategy playing against your opponent
- Samples may not even have full information

The Problem

- Several candidate strategies to choose from
- Only have samples of one strategy playing against your opponent
- Samples may not even have full information
- Problem 1: How can we estimate the performance of the other strategies, based on these samples?

The Problem

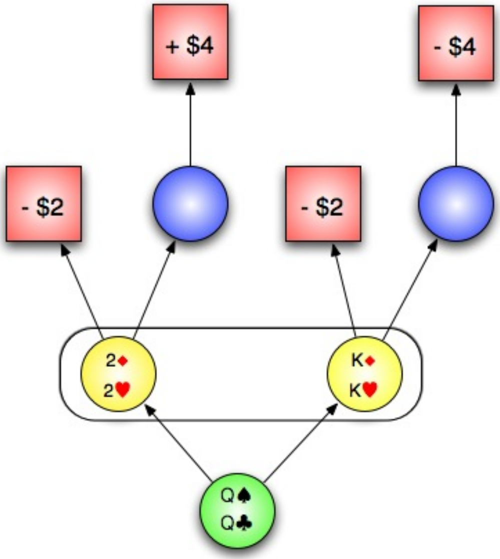
- Several candidate strategies to choose from
- Only have samples of one strategy playing against your opponent
- Samples may not even have full information
- Problem 1: How can we estimate the performance of the other strategies, based on these samples?
- Problem 2: How can we reduce luck (variance) in our estimates?
  - Money = Skill + Luck + Position

The Solution

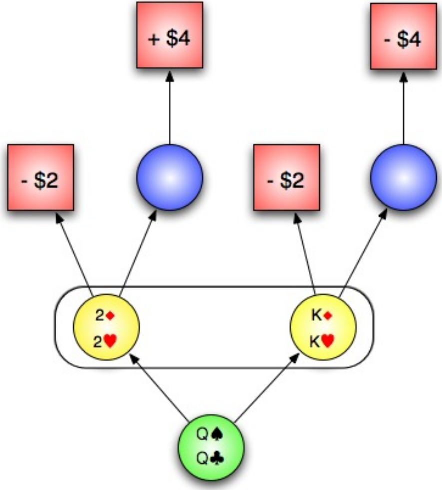
- Importance Sampling for evaluating other strategies
- Combine with existing estimators to reduce variance
- Create additional synthetic data (Main contribution)
- Assumes that the opponent's strategy is static
- General approach, not poker specific

	On Policy	Off Policy
Perfect Information	Unbiased	Bias
Partial Information	Bias	Bias

Repeated Extensive Form Games

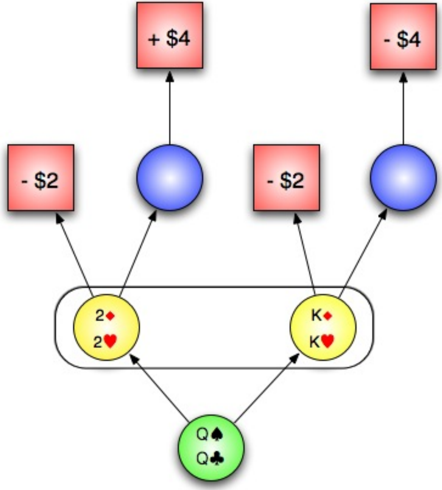


Extensive Form Games



- σ_i - A strategy.
  Action probabilities for player i
- σ - A strategy profile.
  Strategy for each player

Extensive Form Games



- σ_i - A strategy.
  Action probabilities for player i
- σ - A strategy profile.
  Strategy for each player
- $\pi^\sigma(h)$ -Probability of σ reaching h
- $\pi_i^\sigma(h)$ - i's contribution to $\pi^\sigma(h)$
- $\pi_{-i}^\sigma(h)$ - Everyone but i's contribution to $\pi^\sigma(h)$

Importance Sampling

For the terminal nodes $z \in Z$, we can evaluate strategy profile σ with Monte Carlo estimation:

$$E_{z|\sigma}[V(z)] = \frac{1}{t}\sum_{i=1}^{t} V(z_i) \tag{1}$$

- Importance Sampling is a well known technique for estimating the value of one distribution by drawing samples from another distribution
- Useful if one distribution is "expensive" to draw samples from

Importance Sampling for Strategy Evaluation

- σ - strategy profile containing a strategy we want to evaluate
- $\hat{\sigma}$ - strategy profile containing an observed strategy
- In the on-policy case, $\sigma = \hat{\sigma}$

$$E_{z|\hat{\sigma}}[V(z)] = \frac{1}{t}\sum_{i=1}^{t} V(z_i)\frac{\pi^{\sigma}(z)}{\pi^{\hat{\sigma}}(z)} \tag{2}$$

$$= \frac{1}{t}\sum_{i=1}^{t} V(z_i)\frac{\pi_i^{\sigma}(z)\pi_{-i}^{\sigma}(z)}{\pi_i^{\hat{\sigma}}(z)\pi_{-i}^{\hat{\sigma}}(z)} \tag{3}$$

$$= \frac{1}{t}\sum_{i=1}^{t} V(z_i)\frac{\pi_i^{\sigma}(z)}{\pi_i^{\hat{\sigma}}(z)} \tag{4}$$

Importance Sampling for Strategy Evaluation

- σ - strategy profile containing a strategy we want to evaluate
- $\hat{\sigma}$ - strategy profile containing an observed strategy
- In the on-policy case, $\sigma = \hat{\sigma}$

$$E_{z|\hat{\sigma}}[V(z)] = \frac{1}{t}\sum_{i=1}^{t} V(z_i)\frac{\pi^{\sigma}(z)}{\pi^{\hat{\sigma}}(z)} \quad (2)$$

$$= \frac{1}{t}\sum_{i=1}^{t} V(z_i)\frac{\pi_i^{\sigma}(z)\pi_{-i}^{\sigma}(z)}{\pi_i^{\hat{\sigma}}(z)\pi_{-i}^{\hat{\sigma}}(z)} \quad (3)$$

$$= \frac{1}{t}\sum_{i=1}^{t} V(z_i)\frac{\pi_i^{\sigma}(z)}{\pi_i^{\hat{\sigma}}(z)} \quad (4)$$

- Note that the probabilities that depend on the opponent and chance players cancel out!

Basic Importance Sampling and alternate estimators

- On-policy basic importance sampling: just monte-carlo sampling
- Off-policy basic importance sampling: high variance, some bias

Basic Importance Sampling and alternate estimators

- On-policy basic importance sampling: just monte-carlo sampling
- Off-policy basic importance sampling: high variance, some bias
- Any value function can be used
  - For example - the DIVAT estimator for Poker, which is unbiased and low variance

Basic Importance Sampling and alternate estimators

- On-policy basic importance sampling: just monte-carlo sampling
- Off-policy basic importance sampling: high variance, some bias
- Any value function can be used
  - For example - the DIVAT estimator for Poker, which is unbiased and low variance
- We can also create synthetic data. This is the main contribution of the paper.

$U(z')$ and $U^{-1}(z)$

- After observing some terminal histories, you can pretend that something else had happened.

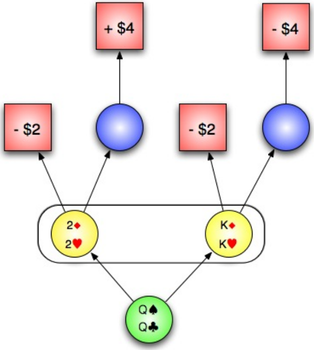
$U(z')$ and $U^{-1}(z)$

- After observing some terminal histories, you can pretend that something else had happened.
- Z is the set of terminal histories
- If we see z, $U^{-1}(z) \subseteq Z$ is the set of synthetic histories we can also evaluate
- Equivalently, if we see a member of $U(z')$, we can also evaluate z'

$U(z')$ and $U^{-1}(z)$

- After observing some terminal histories, you can pretend that something else had happened.
- Z is the set of terminal histories
- If we see z, $U^{-1}(z) \subseteq Z$ is the set of synthetic histories we can also evaluate
- Equivalently, if we see a member of $U(z')$, we can also evaluate z'
- If we choose U carefully, we can still cancel out the opponent's probabilities!
- Two examples - Game-Ending Actions and Other Private Information

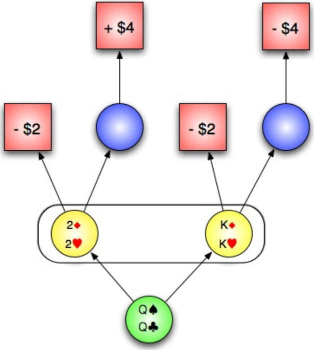
Game-Ending Actions



- h is an observed history

(5)

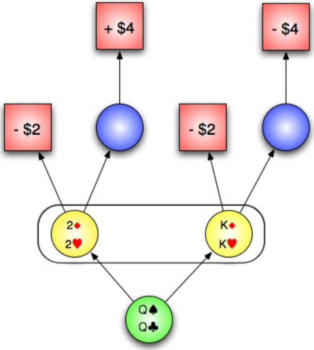
Game-Ending Actions



- h is an observed history
- $S_{-i}(z') \in H$ is a place we could have ended the game

$$ \tag{5} $$

Game-Ending Actions

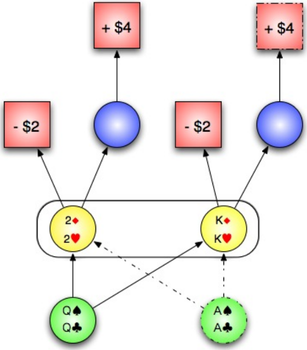


- h is an observed history
- $S_{-i}(z') \in H$ is a place we could have ended the game
- $z' \in U^{-1}(z)$ is the set of synthetic histories where we do end the game

$$\sum_{z' \in U^{-1}(z)} V(z') \frac{\pi_i^{\sigma}(z')}{\pi_i^{\hat{\sigma}}(S_{-i}(z'))} = E_{z|\hat{\sigma}}\left[V(z)\right] \tag{5}$$

Provably unbiased in the on-policy, full information case

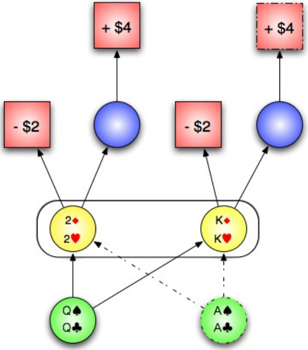
Private Information



- Pretend you had other private information than you actually received
- Opponent's strategy can't depend on our private information

(6)

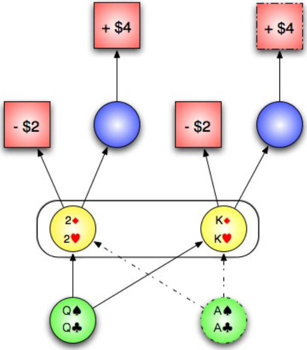
Private Information



- Pretend you had other private information than you actually received
- Opponent's strategy can't depend on our private information
- In poker, pretend you held different 'hole cards'. 2375 more samples per game!

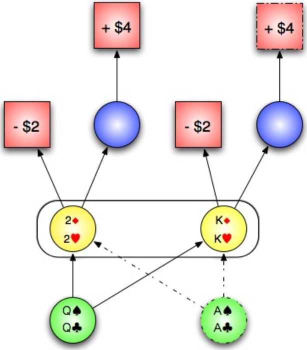
(6)

Private Information



- Pretend you had other private information than you actually received
- Opponent's strategy can't depend on our private information
- In poker, pretend you held different 'hole cards'. 2375 more samples per game!
- $U(z) = \{z' \in Z : \forall \sigma\ \pi^{\sigma}_{-i}(z') = \pi^{\sigma}_{-i}(z)\}$ (6)

Private Information



- Pretend you had other private information than you actually received
- Opponent's strategy can't depend on our private information
- In poker, pretend you held different 'hole cards'. 2375 more samples per game!
- $U(z) = \{z' \in Z : \forall \sigma \; \pi^{\sigma}_{-i}(z') = \pi^{\sigma}_{-i}(z)\}$

$$\sum_{z' \in U^{-1}(z)} V(z') \frac{\pi_i^{\sigma}(z')}{\pi_i^{\hat{\sigma}}(U(z'))} = E_{z|\hat{\sigma}}[V(z)] \tag{6}$$

Provably unbiased in on-policy, full information case

Results

	Bias	StdDev	RMSE
On-Policy: S2298			
Basic	0*	5103	161
BC-DIVAT	0*	2891	91
Game Ending Actions	0*	5126	162
Private Information	0*	4213	133
PI+BC-DIVAT	0*	2146	68
PI+GEA+BC-DIVAT	0*	1778	56
Off-Policy: CFR8			
Basic	200 ± 122	62543	1988
BC-DIVAT	84 ± 45	22303	710
Game Ending Actions	123 ± 120	61481	1948
Private Information	12 ± 16	8518	270
PI+BC-DIVAT	35 ± 13	3254	109
PI+GEA+BC-DIVAT	2 ± 12	2514	80

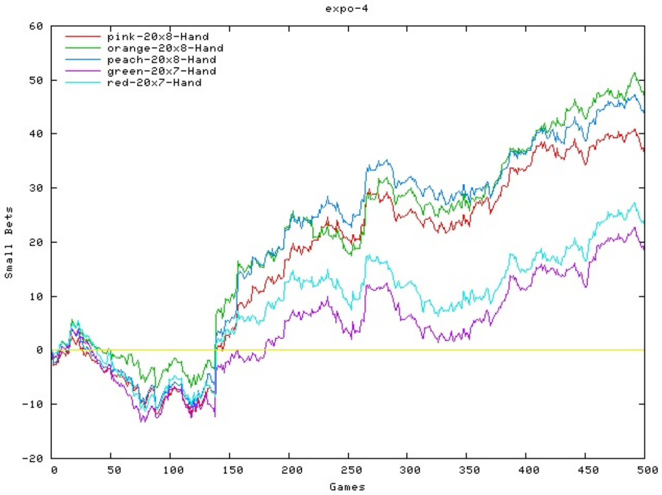
- 1 million hands of S2298 vs PsOpti4
- Units: millibets/game
- RMSE is Root Mean Squared Error over 500 games

Results

	Bias			StdDev			RMSE		
	Min	–	Max	Min	–	Max	Min	–	Max
On Policy									
Basic	0*	–	0*	5102	–	5385	161	–	170
BC-DIVAT	0*	–	0*	2891	–	2930	91	–	92
PI+GEA+BC-DIVAT	0*	–	0*	1701	–	1778	54	–	56
Off Policy									
Basic	49	–	200	20559	–	244469	669	–	7732
BC-DIVAT	10	–	103	12862	–	173715	419	–	5493
PI+GEA+BC-DIVAT	2	–	9	1816	–	2857	58	–	90

- 1 million hands of S2298, CFR8, Orange against PsOpti4
- Units: millibets/game
- RMSE is Root Mean Squared Error over 500 games

Conclusion: Man Machine Poker Championship



Highest Standard Deviation: 1228 millibets/game