Strategy Evaluation in Extensive Games with Importance Sampling

Michael Bowling, Michael Johanson, Neil Burch, Duane Szafron

July 8, 2008
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

<table>
<thead>
<tr>
<th></th>
<th>On Policy</th>
<th>Off Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect Information</td>
<td>Unbiased</td>
<td>Bias</td>
</tr>
<tr>
<td>Partial Information</td>
<td>Bias</td>
<td>Bias</td>
</tr>
</tbody>
</table>
Repeated Extensive Form Games
Extensive Form Games

- $\sigma_i$ - A strategy. Action probabilities for player $i$
- $\sigma$ - A strategy profile. Strategy for each player
Extensive Form Games

- $\sigma_i$ - A strategy. Action probabilities for player $i$
- $\sigma$ - A strategy profile. Strategy for each player
- $\pi^\sigma(h)$ - Probability of $\sigma$ reaching $h$
- $\pi^\sigma_i(h)$ - $i$’s contribution to $\pi^\sigma(h)$
- $\pi^\sigma_{-i}(h)$ - Everyone but $i$’s contribution to $\pi^\sigma(h)$
Importance Sampling

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

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

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

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

Note that the probabilities that depend on the opponent and chance players cancel out!
σ - strategy profile containing a strategy we want to evaluate

ˆσ - strategy profile containing an observed strategy

In the on-policy case, \( \sigma = \hat{\sigma} \)

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

(2)

\[
= \frac{1}{t} \sum_{i=1}^{t} V(z_i) \frac{\pi_{i\sigma}(z)\pi_{i\sigma}(z)}{\pi_{i\hat{\sigma}}(z)\pi_{i\hat{\sigma}}(z)} 
\]

(3)

\[
= \frac{1}{t} \sum_{i=1}^{t} V(z_i) \frac{\pi_{i\sigma}(z)}{\pi_{i\hat{\sigma}}(z)} 
\]

(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.
After observing some terminal histories, you can pretend that something else had happened.
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 \( z' \)
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

- $2$  
- $2$  
- $2$  
+ $4$  
- $4$  

$S^{-i}(z') \in H$ 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

$\sum_{z' \in U^{-1}(z)} V(z') \pi \hat{\sigma}_i(z') \pi \sigma_i(S^{-i}(z')) = E_{z|\hat{\sigma}}[V(z)] \quad (5)$

- $h$ is an observed history
Game-Ending Actions

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

(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

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

Provably unbiased in the on-policy, full information case
Pretend you had other private information than you actually received

Opponent’s strategy can’t depend on our private information

\[
\begin{align*}
\mathbb{U}(z) &= \{ z' \in \mathbb{Z} : \forall \sigma \pi \sigma - i (z') = \pi \sigma - i (z) \} \\
\sum_{z' \in \mathbb{U} - 1}(z') V(z') \pi \sigma i(z') &\approx E_{z | \hat{\sigma}} [V(z)] 
\end{align*}
\]

Provably unbiased in on-policy, full information case
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!
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!

\[ U(z) = \{ z' \in Z : \forall \sigma \pi_\sigma(z') = \pi_{\hat{\sigma}}(z) \} \]

(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!

$U(z) = \left\{ z' \in Z : \forall \sigma \; \pi_\sigma^\sigma_i(z') = \pi^\sigma_i(z) \right\}$

$$\sum_{z' \in U^{-1}(z)} V(z') \frac{\pi_\sigma^\sigma_i(z')}{\pi_\hat{\sigma}_i(U(z'))} = E_{z|\hat{\sigma}} [V(z)]$$

Provably unbiased in on-policy, full information case
<table>
<thead>
<tr>
<th></th>
<th>Bias</th>
<th>StdDev</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>On-Policy: S2298</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td>0*</td>
<td>5103</td>
<td>161</td>
</tr>
<tr>
<td>BC-DIVAT</td>
<td>0*</td>
<td>2891</td>
<td>91</td>
</tr>
<tr>
<td>Game Ending Actions</td>
<td>0*</td>
<td>5126</td>
<td>162</td>
</tr>
<tr>
<td>Private Information</td>
<td>0*</td>
<td>4213</td>
<td>133</td>
</tr>
<tr>
<td>PI+BC-DIVAT</td>
<td>0*</td>
<td>2146</td>
<td>68</td>
</tr>
<tr>
<td>PI+GEA+BC-DIVAT</td>
<td>0*</td>
<td>1778</td>
<td>56</td>
</tr>
<tr>
<td><strong>Off-Policy: CFR8</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td>200 ± 122</td>
<td>62543</td>
<td>1988</td>
</tr>
<tr>
<td>BC-DIVAT</td>
<td>84 ± 45</td>
<td>22303</td>
<td>710</td>
</tr>
<tr>
<td>Game Ending Actions</td>
<td>123 ± 120</td>
<td>61481</td>
<td>1948</td>
</tr>
<tr>
<td>Private Information</td>
<td>12 ± 16</td>
<td>8518</td>
<td>270</td>
</tr>
<tr>
<td>PI+BC-DIVAT</td>
<td>35 ± 13</td>
<td>3254</td>
<td>109</td>
</tr>
<tr>
<td>PI+GEA+BC-DIVAT</td>
<td>2 ± 12</td>
<td>2514</td>
<td>80</td>
</tr>
</tbody>
</table>

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

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

- 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