Efficient Probabilistic Performance Bounds for Inverse Reinforcement Learning

Daniel Brown and Scott Niekum
The University of Texas at Austin
Learning from Demonstration (LfD)
Bounding Performance for LfD

- Correctness
- Generalizability
- Safety
Bounding Policy Loss

- Value of policy

\[ V_\pi^R = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t) \right] \]

- Policy Loss

\[ V_\pi^{\ast} - V_\pi^R \]
General Problem: Policy evaluation w/out R

- Given:
  - Domain, MDP\(\text{MDP}\backslash R\)
  - Demonstrations, \(D\)
  - Evaluation policy, \(\pi_{\text{eval}}\)

- Find \(\epsilon\)
  such that with high confidence

\[
V_R^{\pi^*} - V_R^{\pi_{\text{eval}}} \leq \epsilon
\]
How to bound Policy Loss?

\[ V^\pi_* - V^\pi_{eval} \leq \epsilon \]

- We don’t know the reward function (or the optimal policy)
  - Bayesian Inverse Reinforcement Learning
Bayesian IRL (Ramachandran 2007)

• Uses MCMC to sample from posterior

\[ P(R|D) \propto P(D|R)P(R) \]

• Assumes demonstrations follow softmax policy with temperature \( c \).

\[ P(D|R) = \prod_{(s,a) \in D} \frac{e^{cQ^*_R(s,a)}}{\sum_{b \in A} e^{cQ^*_R(s,b)}} \]
How to bound Policy Loss?

\[ V^{\pi^*}_R - V^{\pi_{\text{eval}}}_R \leq \epsilon \]

- We don’t know the reward function (or the optimal policy)
  - Bayesian Inverse Reinforcement Learning
How to bound Policy Loss?

\[ V_{R}^{\pi^*} - V_{R}^{\pi_{eval}} \leq \epsilon \]

- We don’t know the reward function (or the optimal policy)
  - Bayesian Inverse Reinforcement Learning
  - Risk-sensitive performance bound
    - \( \alpha \)-Value at Risk (\( \alpha \)-quantile worst-case outcome)
High-level Approach

\[ \pi_{\text{eval}} \]

\[ \mathcal{D} \]

\[ \mathcal{C} \]

\[ P(R) \]

\[ P(R|D) \]

\[ V_{\pi^*_{R_i}} - V_{\pi_{\text{eval}}_{R_i}} \]

Bayesian IRL

Calculate policy loss assuming sampled reward is true reward

Sorted Policy Losses

Return high confidence bound on alpha-worst-case policy loss over \( P(R|D) \).
Experiments

- Grid world
- Driving
Assumptions on Reward Functions

- Linear combination of features
  \[ R(s) = w^T \phi(s) \quad \| w \|_1 \leq 1 \]

- We can rewrite the expected return of a policy in terms of expected feature counts
  \[ V_R^\pi = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t w^T \phi(s_t) | \pi \right] = w^T \mu(\pi) \]
Baseline

• Worst-case feature count bound (WFCB)
  – Penalize the largest difference in state-visitation counts between demonstrations and evaluation policy

\[
\text{WFCB}(\pi_{\text{eval}}, D) = \left\| \hat{\mu}^* - \mu(\pi_{\text{eval}}) \right\|_{\infty}
\]
Grid World Results

• 200 random grid worlds.

• Evaluation policy is optimal policy for MAP reward given demonstrations

![Diagram showing efficiency gain with increasing number of demonstrations]
Theoretical IRL performance bounds

• Based on Hoeffding-style concentration inequalities
  – (Abbeel & Ng 2004, Syed & Schapire 2008)
• Extremely loose in practice

<table>
<thead>
<tr>
<th>Number of demonstrations</th>
<th>Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.98</td>
</tr>
<tr>
<td>9</td>
<td>0.9372</td>
</tr>
<tr>
<td>23,146</td>
<td>0.1328</td>
</tr>
<tr>
<td>0.95-VaR Bound</td>
<td>0.98</td>
</tr>
<tr>
<td>Syed and Schapire Bound</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>0.9372</td>
</tr>
<tr>
<td></td>
<td>142.59</td>
</tr>
<tr>
<td></td>
<td>47.53</td>
</tr>
</tbody>
</table>
Policy Selection

• Rank a set of evaluation policies based on high-confidence performance bounds
Driving Experiment

• Actions = left, right, straight
• State Features: distances to other cars, lane #
• Reward features: lane #, in collision
Demonstration that avoids collisions

**Right-safe:** avoids cars but prefers right lane

**On-road:** Stays on road, but ignores other cars

**Nasty:** seeks collisions
Policy Ranking

<table>
<thead>
<tr>
<th>$\pi_{eval}$</th>
<th>Collisions</th>
<th>True</th>
<th>WFCB</th>
<th>0.95-VaR</th>
</tr>
</thead>
<tbody>
<tr>
<td>right-safe</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>on-road</td>
<td>13.65</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>nasty</td>
<td>42.75</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

- Feature count bound is misled by state-occupancies
- Our method reasons over reward likelihoods
Future Work

• Scalability:

• Estimating the amount of noise in human demonstrations

• Active Learning: query demonstrator to reduce VaR
Conclusion

• First practical method for policy evaluation when reward function is unknown.

• Based on probabilistic worst-case performance over likely reward functions.

• Applications:
  – Policy selection
  – Policy improvement
  – Demonstration sufficiency
Future Work

• Scalability:

• Estimating the noise in human demonstrations

• Active Learning: query demonstrator to reduce VaR