Feb 2018

TEXAS
The University of Texas at Austin

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

π

- Correctness
- Generalizability
- Safety

Bounding Policy Loss

How close is my performance to optimal?

- Value of policy

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

- Policy Loss

$$V_R^{\pi^*} - V_R^{\pi}$$

General Problem: Policy evaluation w/out R

- Given:
  - Domain, MDP\R
  - Demonstrations, D
  - Evaluation policy, π_{eval}
- Find ϵ
  such that with high confidence

$$V_R^{\pi^*} - V_R^{\pi_{\text{eval}}} \leq \epsilon$$

I'm 95% confident my performance is ε-close to optimal.

How to bound Policy Loss?

$$V_R^{\pi^*} - V_R^{\pi_{\text{eval}}} \leq \epsilon$$

- 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?

$$V_R^{\pi^*} - V_R^{\pi_{\text{eval}}} \leq \epsilon$$

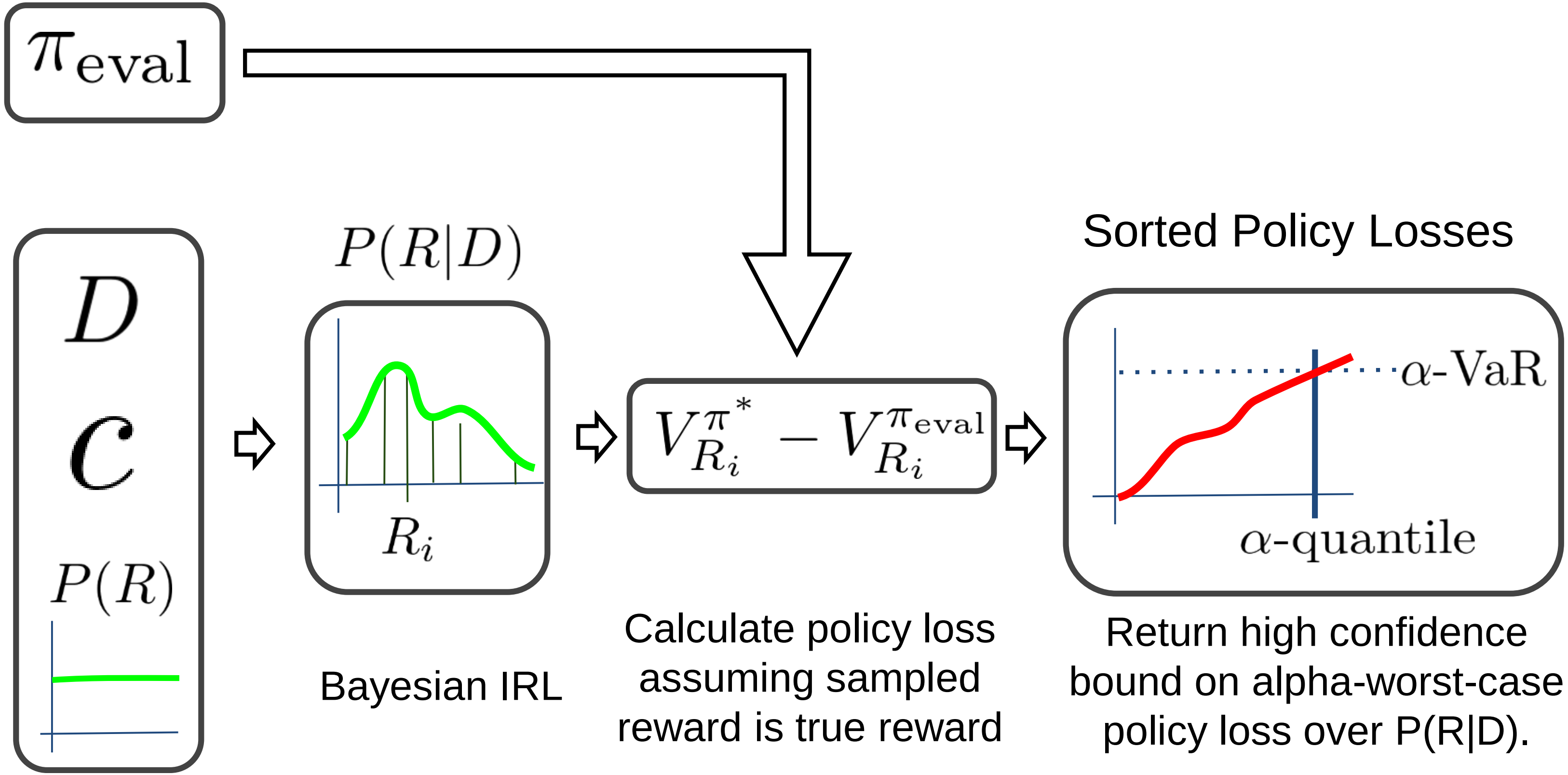
- We don’t know the reward function (or the optimal policy)
  - **Bayesian Inverse Reinforcement Learning**

How to bound Policy Loss?

$$V_R^{\pi^*} - V_R^{\pi_{\text{eval}}} \leq \epsilon$$

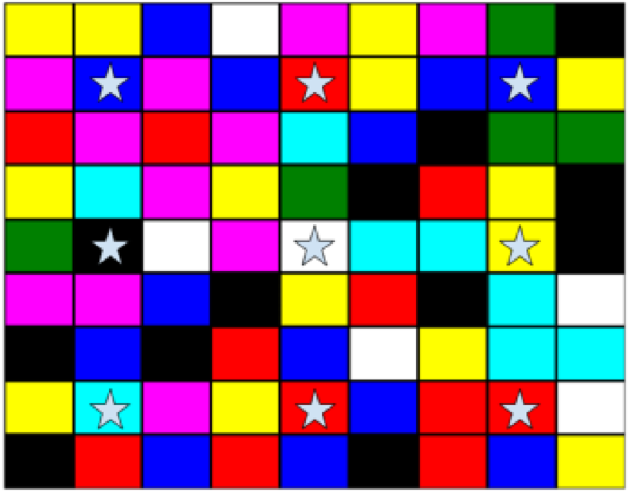
- We don't know the reward function (or the optimal policy)
  - Bayesian Inverse Reinforcement Learning
  - **Risk-sensitive performance bound**
    - α-Value at Risk (α-quantile worst-case outcome)

High-level Approach

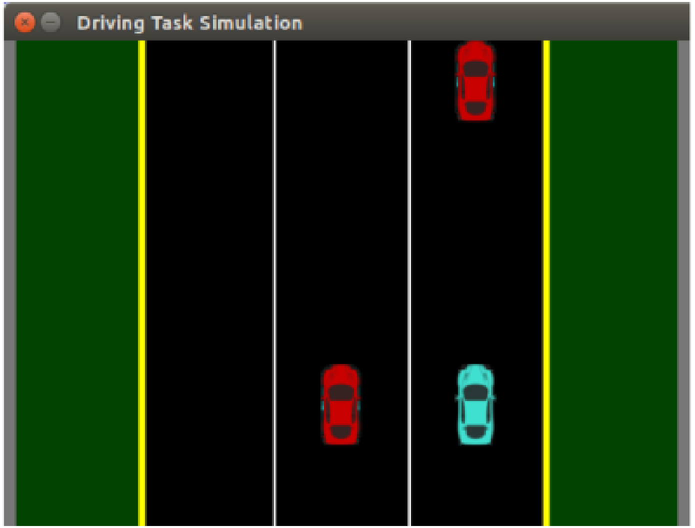


Experiments

- Grid world



- Driving



Assumptions on Reward Functions

- Linear combination of features

$$R(s) = w^T \phi(s) \qquad\qquad \|w\|_1 \leq 1$$

- We can rewrite the expected return of a policy in terms of expected feature counts

$$V_R^\pi = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t w^T \phi(s_t) | \pi] = w^T \mu(\pi)$$

Baseline

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

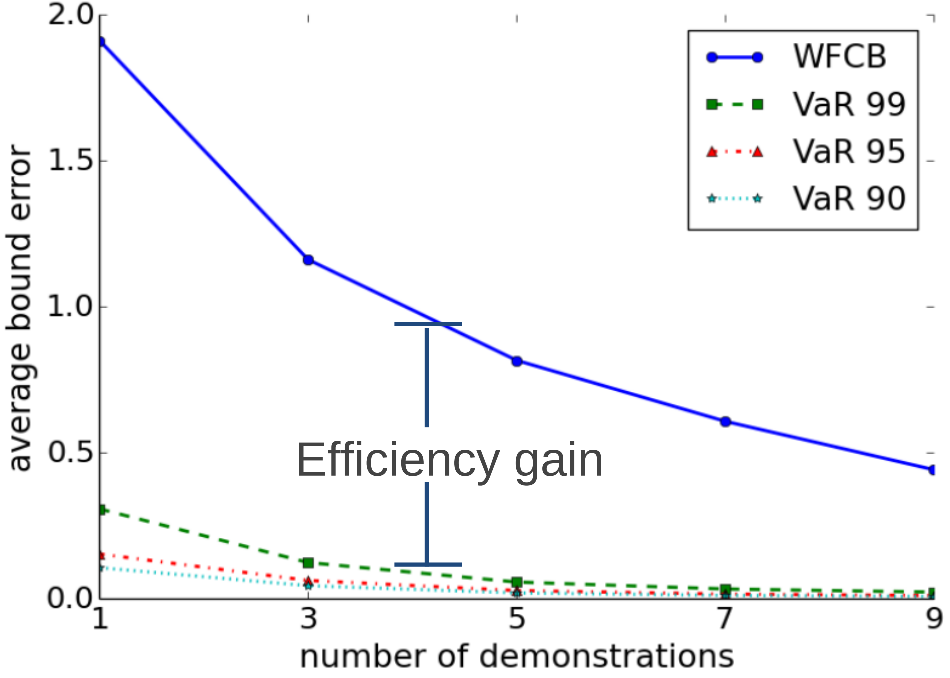
$$\mathrm{WFCB}(\pi_{\mathrm{eval}}, D) = \|\hat{\mu}^* - \mu(\pi_{\mathrm{eval}})\|_\infty$$

Empirical expected feature counts of demonstrations

Expected feature counts of evaluation policy

Grid World Results

- 200 random grid worlds.
- Evaluation policy is optimal policy for MAP reward given demonstrations



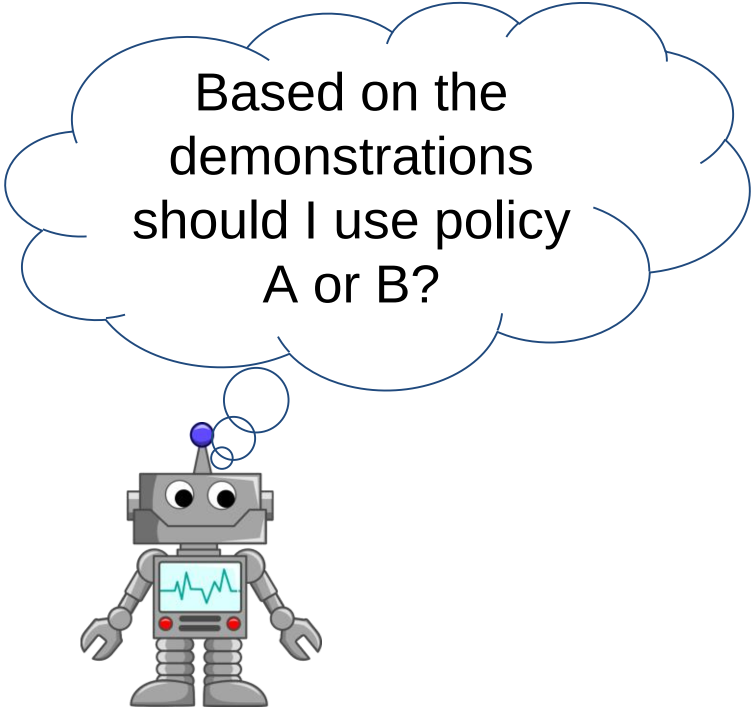
Theoretical IRL performance bounds

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

	Number of demonstrations			Average Accuracy
	1	9	···	23,146	
0.95-VaR Bound	**0.9372**	**0.1328**		-	0.98
Syed and Schapire Bound	142.59	47.53		0.9372	1.0

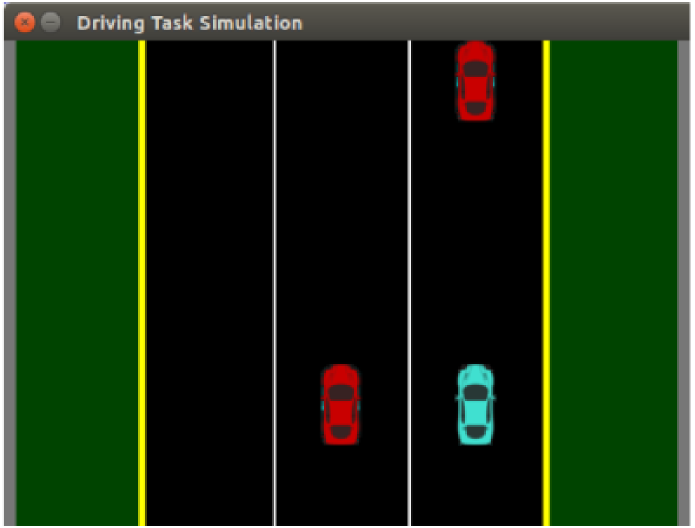
Policy Selection

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

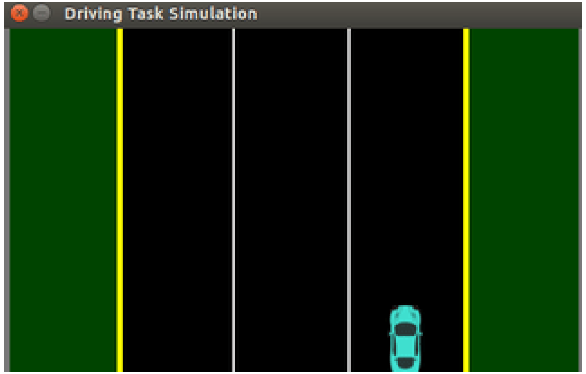


Driving Experiment

- Actions = left, right, straight
- State Features: distances to other cars, lane #
- Reward features: lane #, in collision

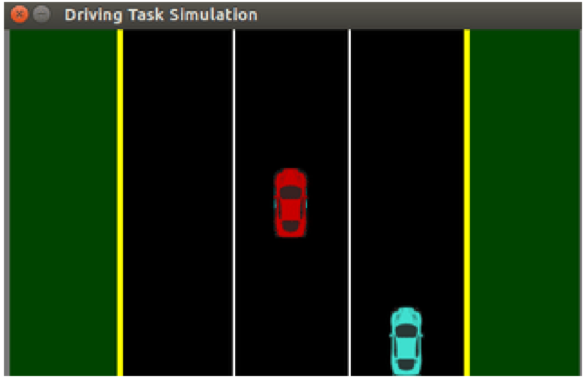


Demonstration that avoids collisions



Right-safe: avoids cars but prefers right lane

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



Nasty: seeks collisions

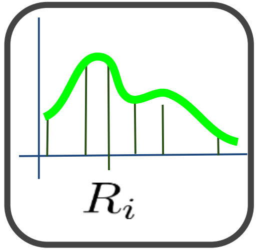
Policy Ranking

		Ranking		
π_{eval}	Collisions	True	WFCB	0.95-VaR
right-safe	0	**1**	3	**1**
on-road	13.65	**2**	1	**2**
nasty	42.75	**3**	2	**3**

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

Future Work

- Scalability:



$$\boxed{V_{R_i}^{\pi^*} - V_{R_i}^{\pi_{\text{eval}}}}$$

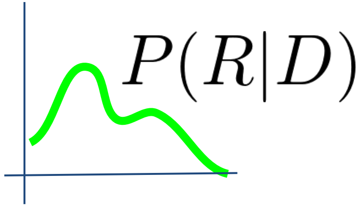
- Estimating the amount of noise in human demonstrations

- Active Learning: query demonstrator to reduce VaR

Conclusion

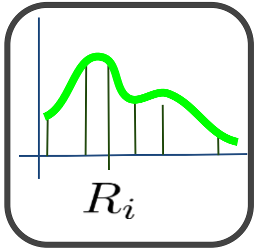
- First practical method for policy evaluation when reward function is unknown.
- Based on probabilistic worst-case performance over likely reward functions.
- Applications:
  - Policy selection
  - Policy improvement
  - Demonstration sufficiency

MDP\R



Future Work

- Scalability:



$$V_{R_i}^{\pi^*} - V_{R_i}^{\pi_{\text{eval}}}$$

- Estimating the noise in human demonstrations

- Active Learning: query demonstrator to reduce VaR