SCALING PROBABILISTICALLY SAFE LEARNING TO ROBOTICS

Scott Niekum

Assistant Professor, Department of Computer Science

The University of Texas at Austin

TEXAS

The University of Texas at Austin

PeARL

Personal Autonomous Robotics Lab

Safety and Correctness in Robotics

What does it mean for a learning agent to be "safe"?

- **Formal safety:** A self-driving car that will provably never crash if some model holds
- **Risk-sensitive safety:** A stock market agent with bounded value-at-risk
- **Robust safety:** An image classifier resistant to data poisoning or adversarial examples
- **Monotonic safety:** An RL-based advertising policy that always improves with high probability
- **Safe exploration:** A walking robot that can explore new gaits without falling over

More complete taxonomy: D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. "Concrete problems in AI safety."

A proposed definition of safety:

Safety = Correctness + Confidence

Correctness: Meeting or exceeding a measure of performance

Confidence: A (probabilistic) guarantee of correctness

A spectrum of safety

Guaranteed ⟵⟶ Probabilistic ⟵⟶ Approximate

Guaranteed

Require perfect models

Verification / synthesis

[Kress-Gazit et. al 2009]
[Raman et. al 2015]

Probabilistic

Sample inefficient

PAC-MDP methods

[Singh et. al 2002]
[Fu and Topcu 2014]

Concentration inequalities

[Thomas et. al 2015]
[Bottou et. al 2013]
[Abbeel and Ng 2004]
[Syed and Schapire 2008]

Approximate

No guarantees

KL-divergence constraints

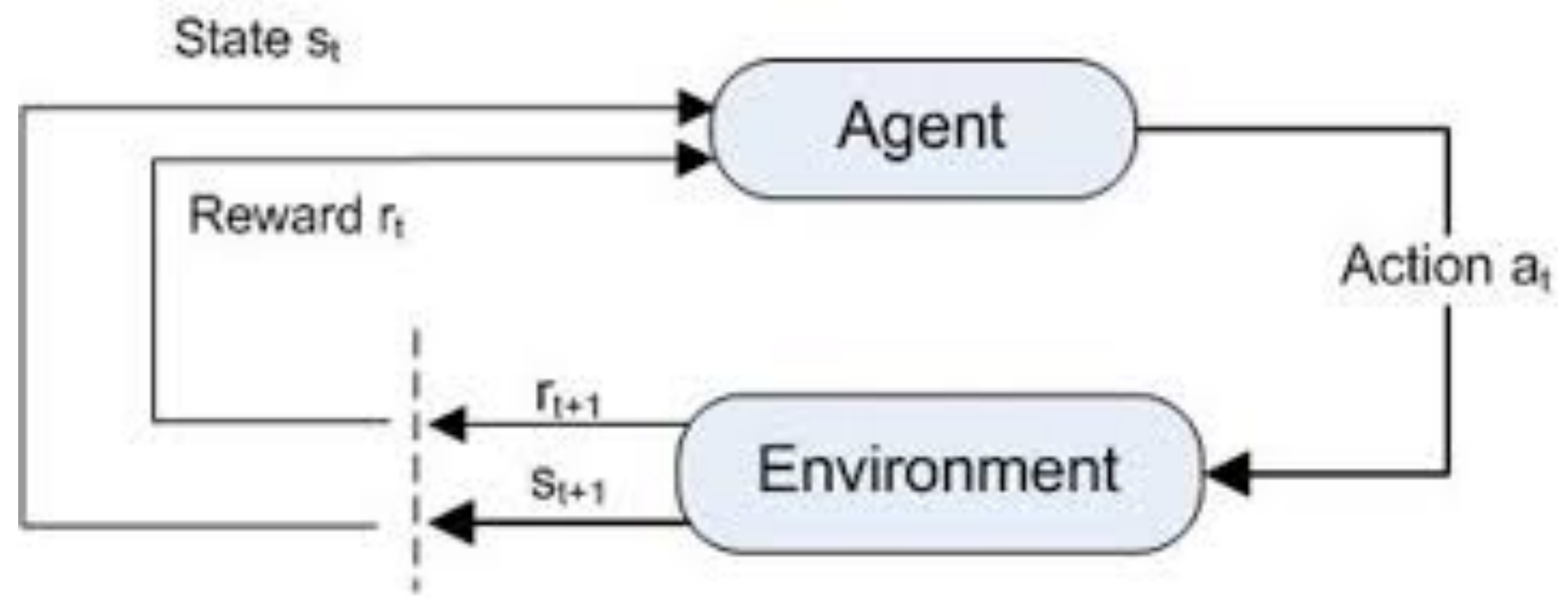
[Schulman et. al 2015]
[Schulman et. al 2017]
[Peters et. al 2010]

Address bad assumptions!

Part 1: Safe reinforcement learning

Part 2: Safe imitation learning

Background



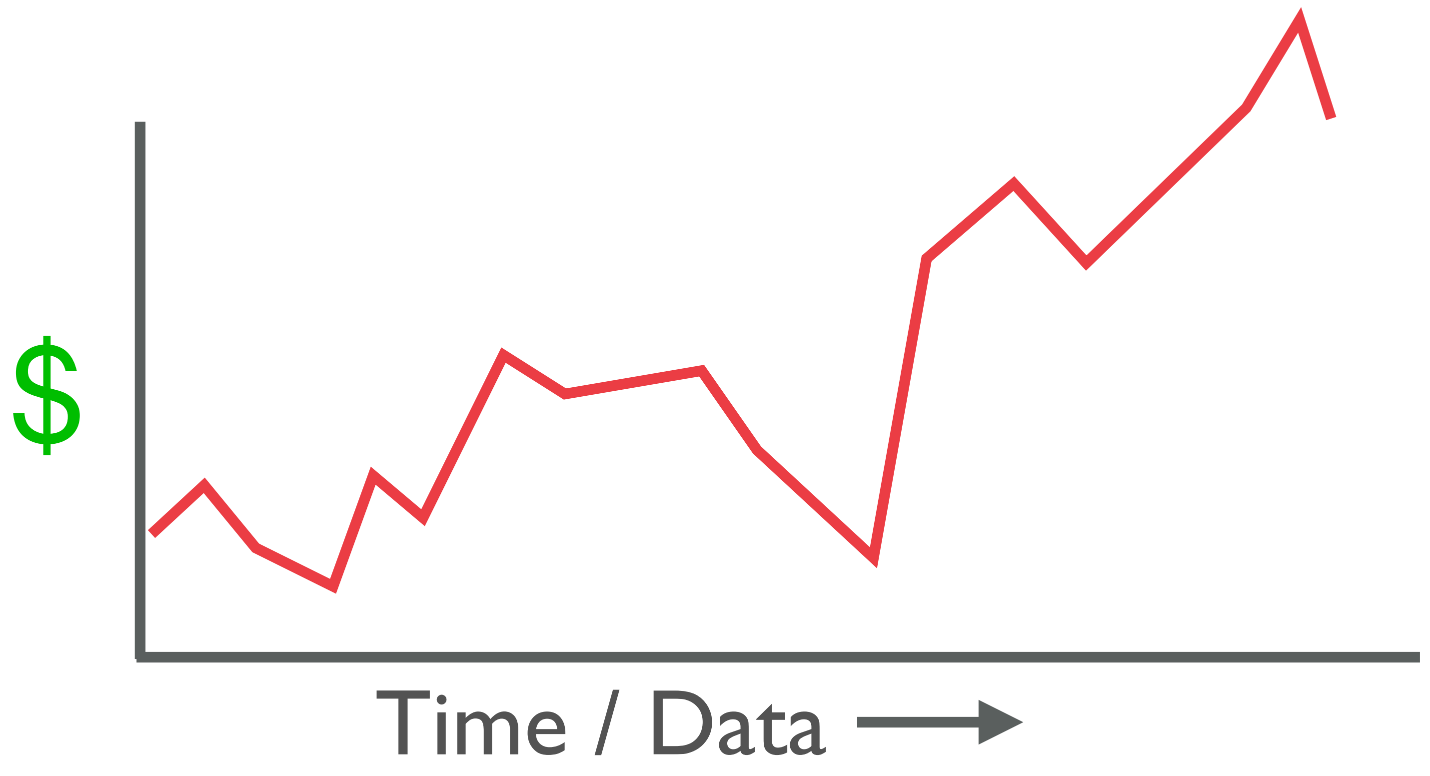
- Finite-horizon MDP.
- Agent selects actions with a *stochastic* policy, π.
- The policy and environment determine a distribution over trajectories, $H : S_0, A_0, R_0, S_1, A_1, R_1, ..., S_L, A_L, R_L$

Safe off-policy evaluation (OPE):

Determine a **probabilistic lower bound** on **expected performance** of a policy, given data generated by a **different policy**

Safe policy improvement (PI):

Ensure that expected performance **improves monotonically** at every learning step with **high confidence**



Policy Evaluation

Policy performance:

$$V(\pi) = \mathbb{E}\left[\sum_{t=0}^{L} \gamma^t R_t \middle| H \sim \pi\right]$$

Given a target policy, π_e, estimate $V(\pi_e)$

- Let $\pi_e \equiv \pi_{\boldsymbol{\theta}_e}$

Monte Carlo Policy Evaluation

Given a dataset $\mathcal{D}$ of trajectories where $\forall H \in \mathcal{D}$, $H \sim \pi_e$:

$$\text{MC}(\mathcal{D}) := \frac{1}{|\mathcal{D}|} \sum_{H_i \in \mathcal{D}} \sum_{t=0}^{L} \gamma^t R_t^{(i)}$$

Importance Sampling Policy Evaluation[1]

Given a dataset $\mathcal{D}$ of trajectories where $\forall H_i \in \mathcal{D}$, H_i is sampled from a behavior policy π_i:

$$\text{IS}(\mathcal{D}) := \frac{1}{|\mathcal{D}|} \sum_{H_i \in \mathcal{D}} \underbrace{\prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi_i(A_t|S_t)}}_{\text{re-weighting factor}} \sum_{t=0}^{L} \gamma^t R_t^{(i)}$$

For convenience:

$$\text{IS}(H, \pi) := \prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi(A_t|S_t)} \sum_{t=0}^{L} \gamma^t R_t$$

[1] Precup, Sutton, and Singh (2000)

Confidence Intervals for Off-Policy Evaluation

Given:

- Trajectories generated by a *behavior* policy, π_b, $\{H, \pi_b\} \in \mathcal{D}$.
- An *evaluation* policy, π_e.
- $\delta \in [0, 1]$ is a confidence level.

Determine a lower bound $\hat{V}_{\text{lb}}(\pi_e, \mathcal{D})$ such that $V(\pi_e) \geq \hat{V}_{\text{lb}}(\pi_e, \mathcal{D})$ with probability $1 - \delta$.

Concentration Inequalities

Chernoff-Hoeffding Inequality

- Probabilistic bound on how a random variable deviates from its expectation
- No distributional assumptions
- With probability at least 1-δ:

$$\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i - b\sqrt{\frac{\log(1/\delta)}{2n}}$$

- Can use with importance sampled returns to bound value of a policy from off-policy samples
- Significantly tighter bounds exist under certain conditions (Thomas et. al 2015)

Sample (in)efficiency (Thomas et. al 2015)

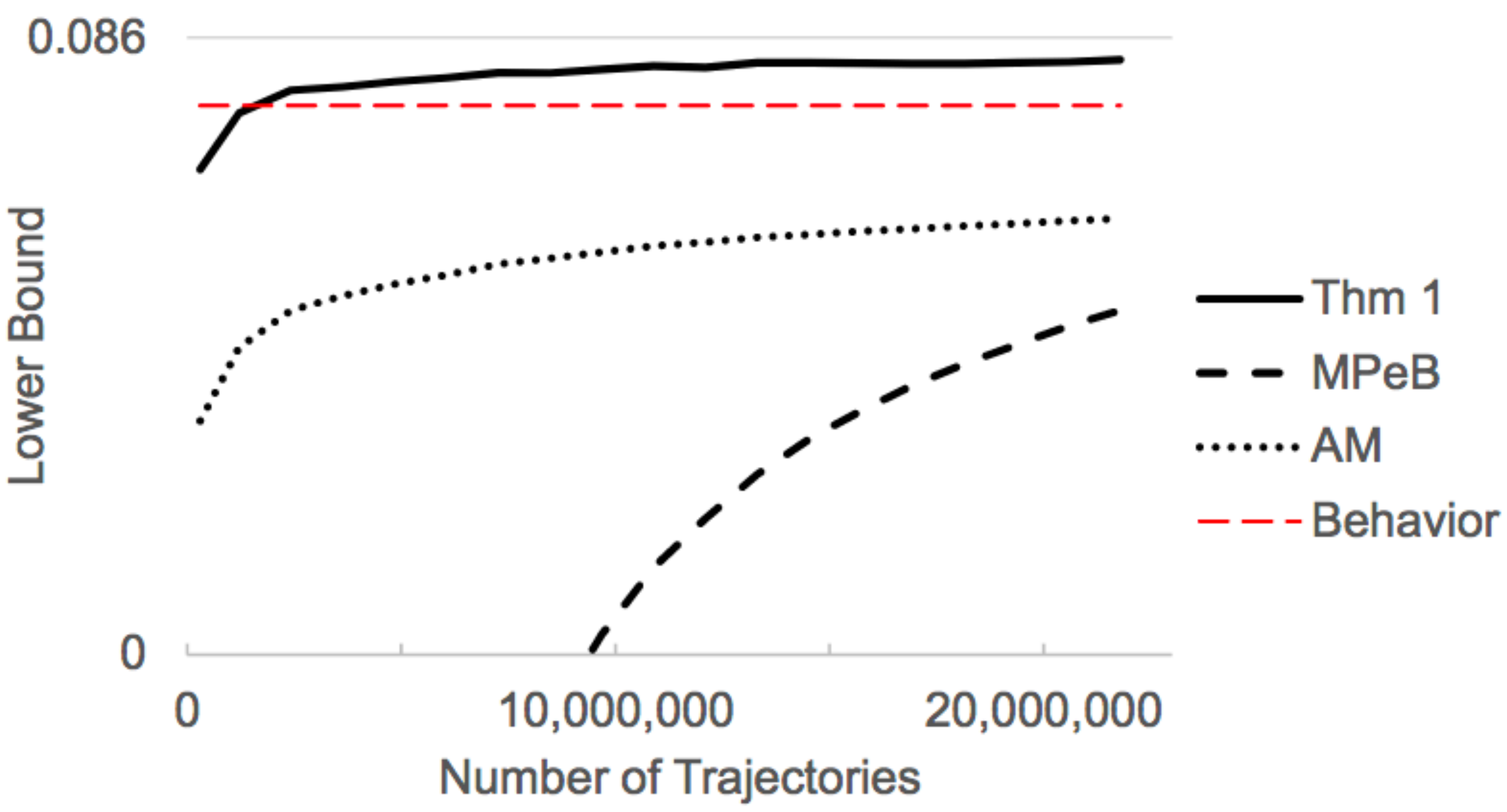


Figure 3: 95% confidence lower bound (unnormalized) on $\rho(\theta)$ using trajectories generated using the simulator described in the text. The behavior policy's true expected re-

Bad assumption #1:

"When performing policy evaluation, it is better to collect on-policy data than off-policy data"

J.P. Hanna, P.S. Thomas, P. Stone, and S. Niekum.
Data-Efficient Policy Evaluation Through Behavior Policy Search.
Proceedings of the 34th International Conference on Machine Learning (ICML), August 2017.

Optimal Behavior Policy

Claim: There exists an optimal behavior policy, $\pi_{b^\star}$, if all returns are positive and transitions are deterministic:

Optimal Behavior Policy

Claim: There exists an optimal behavior policy, $\pi_{b^\star}$, if all returns are positive and transitions are deterministic:

$$V(\pi_e) = g(H) \prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi_{b^\star}(A_t|S_t)}$$

$$\prod_{t=0}^{L} \pi_{b^\star}(A_t|S_t) = \frac{g(H)}{V(\pi_e)} \prod_{t=0}^{L} \pi_e(A_t|S_t)$$

$$w_{\pi_{b^\star}}(H) = \frac{g(H)}{V(\pi_e)} w_{\pi_e}(H)$$

Zero mean squared error with a single trajectory! Such a policy provably exists as a mixture over time-dependent deterministic policies (i.e. weighted trajectories).

Optimal Behavior Policy

Unfortunately, the optimal behavior policy is unknown in practice.

$$\prod_{t=0}^{L} \pi_{b^\star}(A_t|S_t) = \frac{g(H)}{V(\pi_e)} \prod_{t=0}^{L} \pi_e(A_t|S_t)$$

- Requires $V(\pi_e)$ be known!
- Requires the reward function be known.
- Requires deterministic transitions.

Behavior Policy Gradient

Key Idea: Adapt the behavior policy parameters, $\boldsymbol{\theta}$, with gradient descent on the mean squared error of importance-sampling.

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \alpha \frac{\partial}{\partial \boldsymbol{\theta}} \text{MSE}[\text{IS}(H_i, \boldsymbol{\theta})]$$

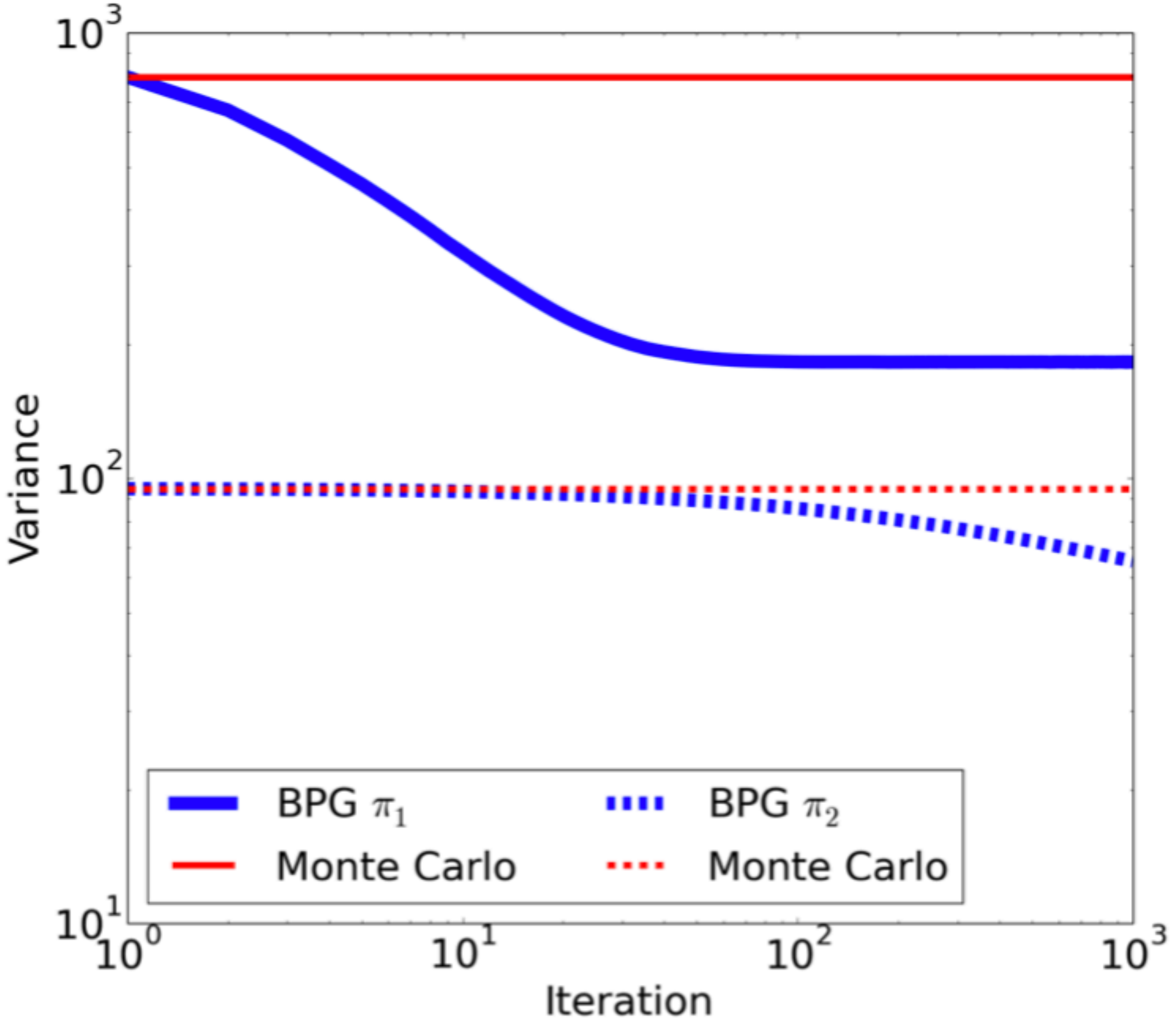
- $\text{MSE}[IS(H, \boldsymbol{\theta})]$ is **not** computable.
- $\frac{\partial}{\partial \boldsymbol{\theta}} \text{MSE}[IS(H, \boldsymbol{\theta})]$ is computable.

Behavior Policy Gradient Theorem

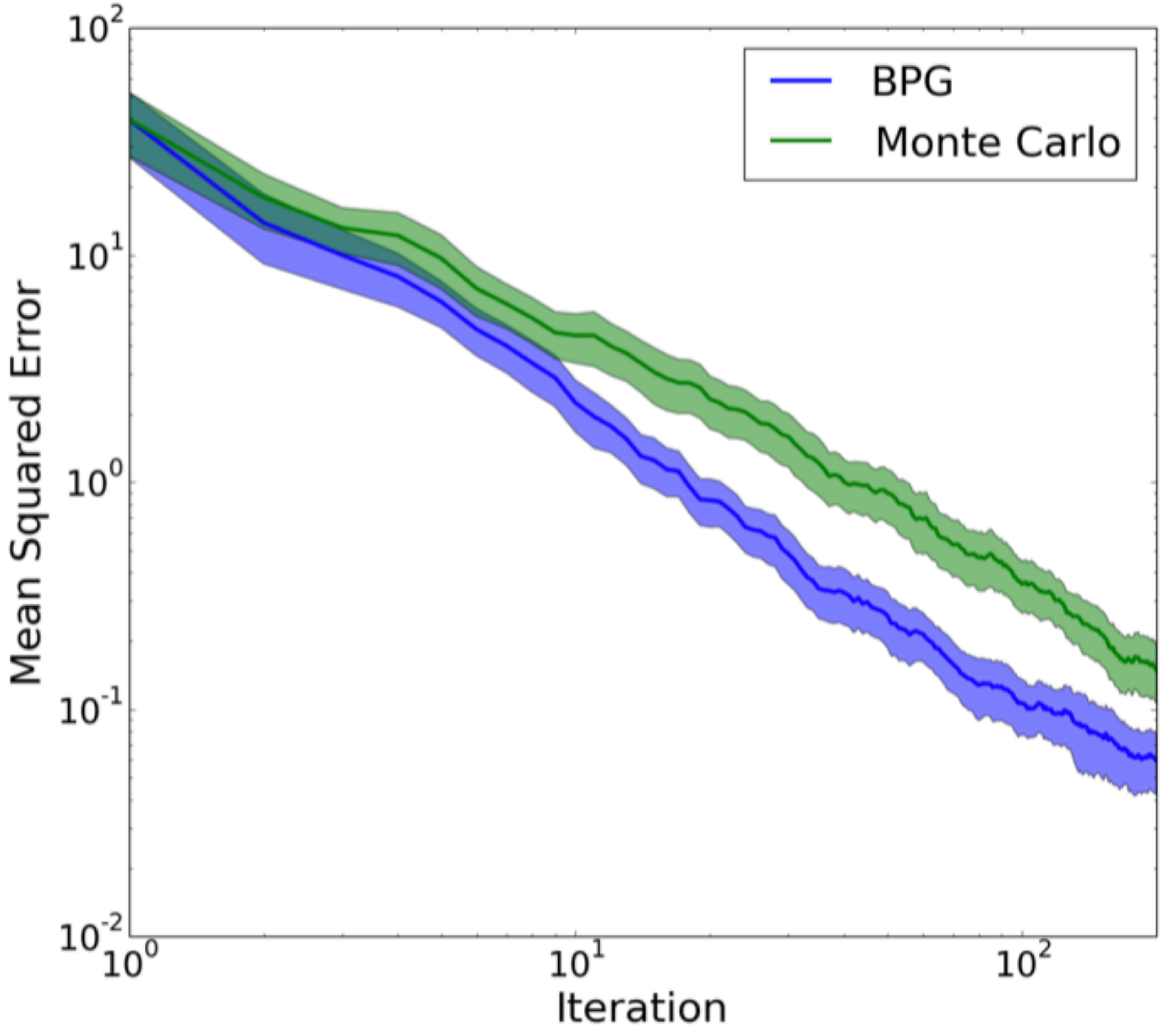
Theorem

$$\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}(\operatorname{IS}(H, \boldsymbol{\theta})) = \mathbf{E}_{\pi_{\boldsymbol{\theta}}}\left[-\operatorname{IS}(H, \boldsymbol{\theta})^2 \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(\pi_{\boldsymbol{\theta}}(A_t | S_t)\right)\right]$$

Variance reduction



Improved sample efficiency

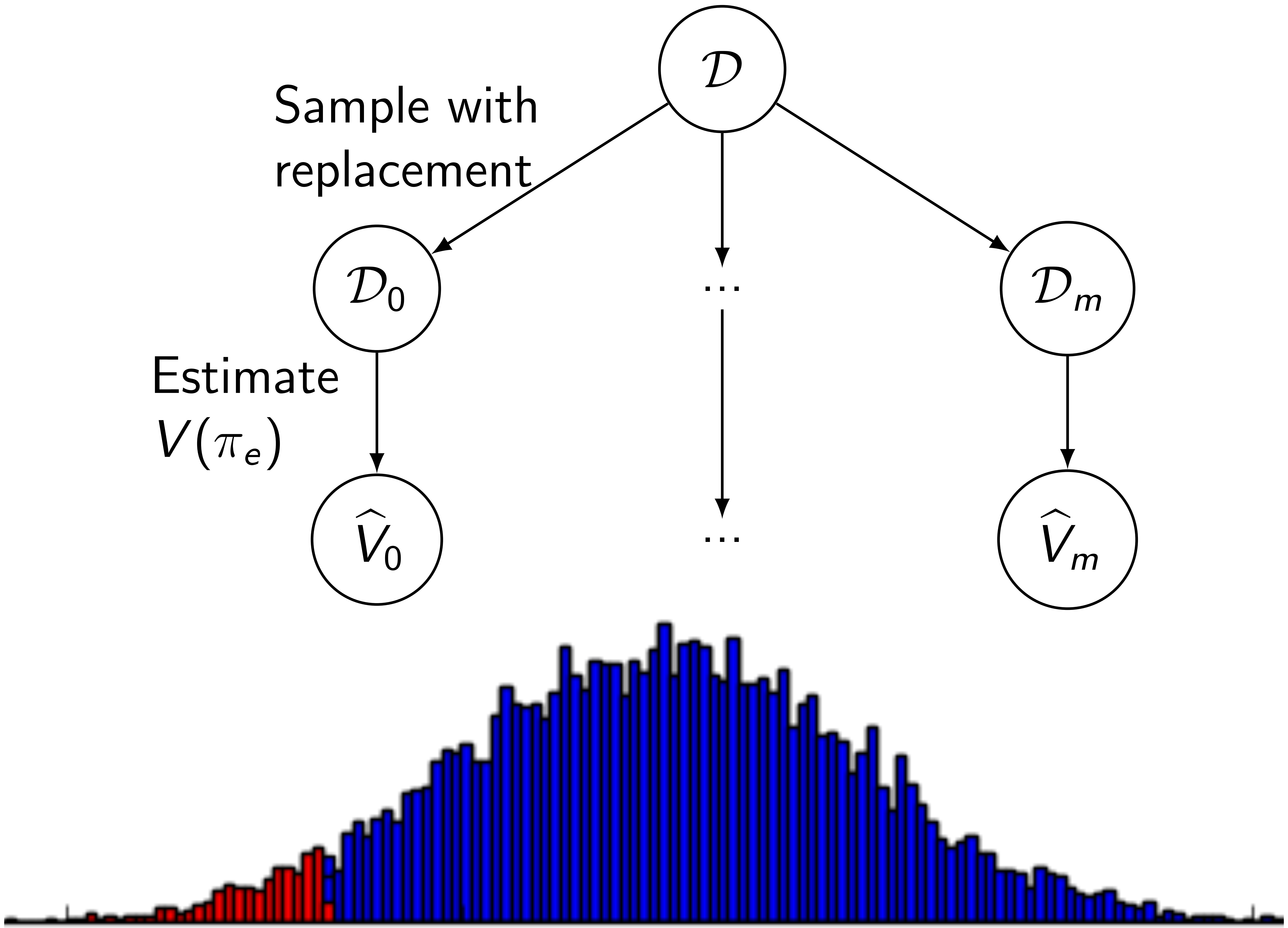


Cartpole Swing-up

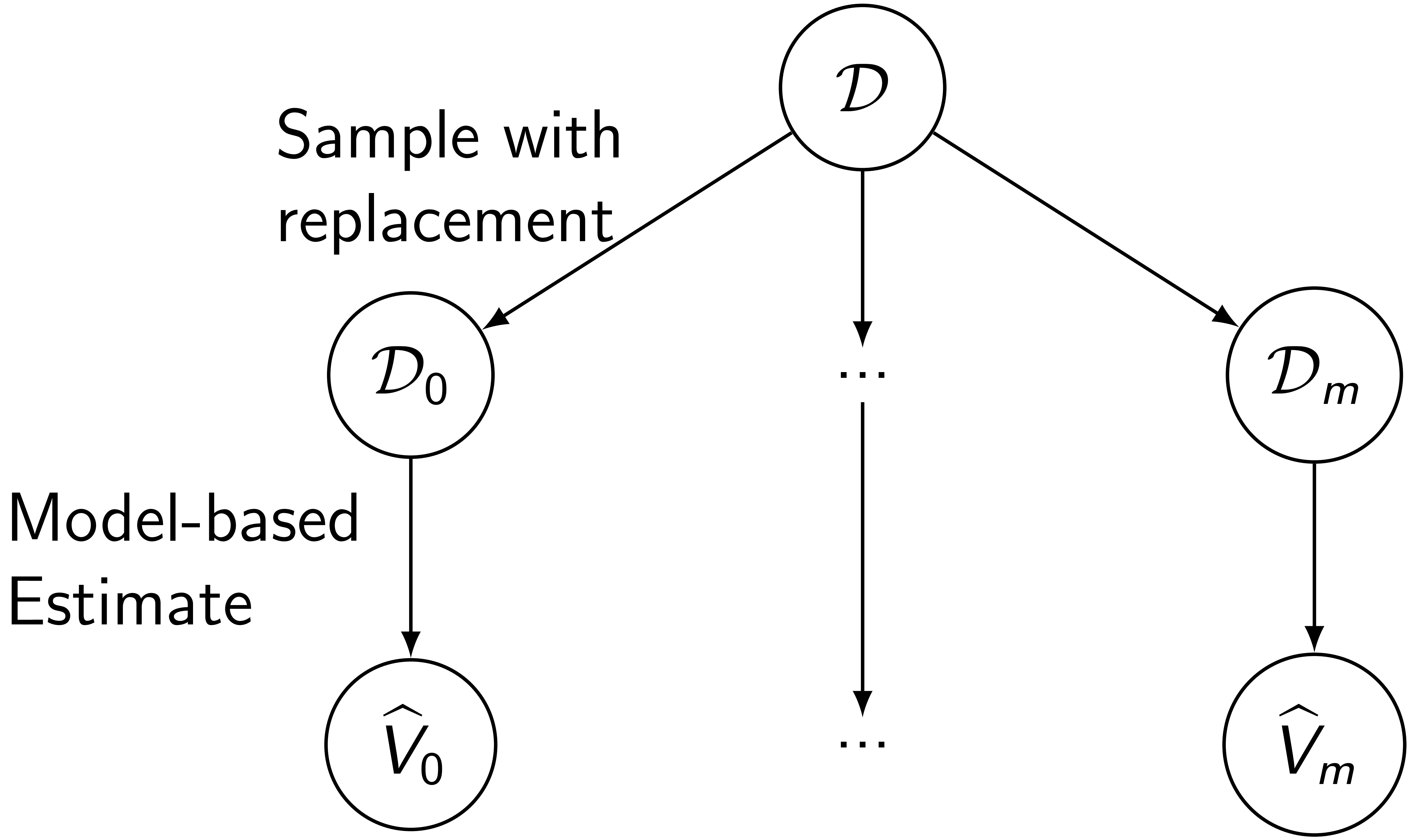
Better, but not good enough.

- Are "semi-safe", consistent methods good enough? (e.g. bootstrapping)
- Why only use model-free methods?

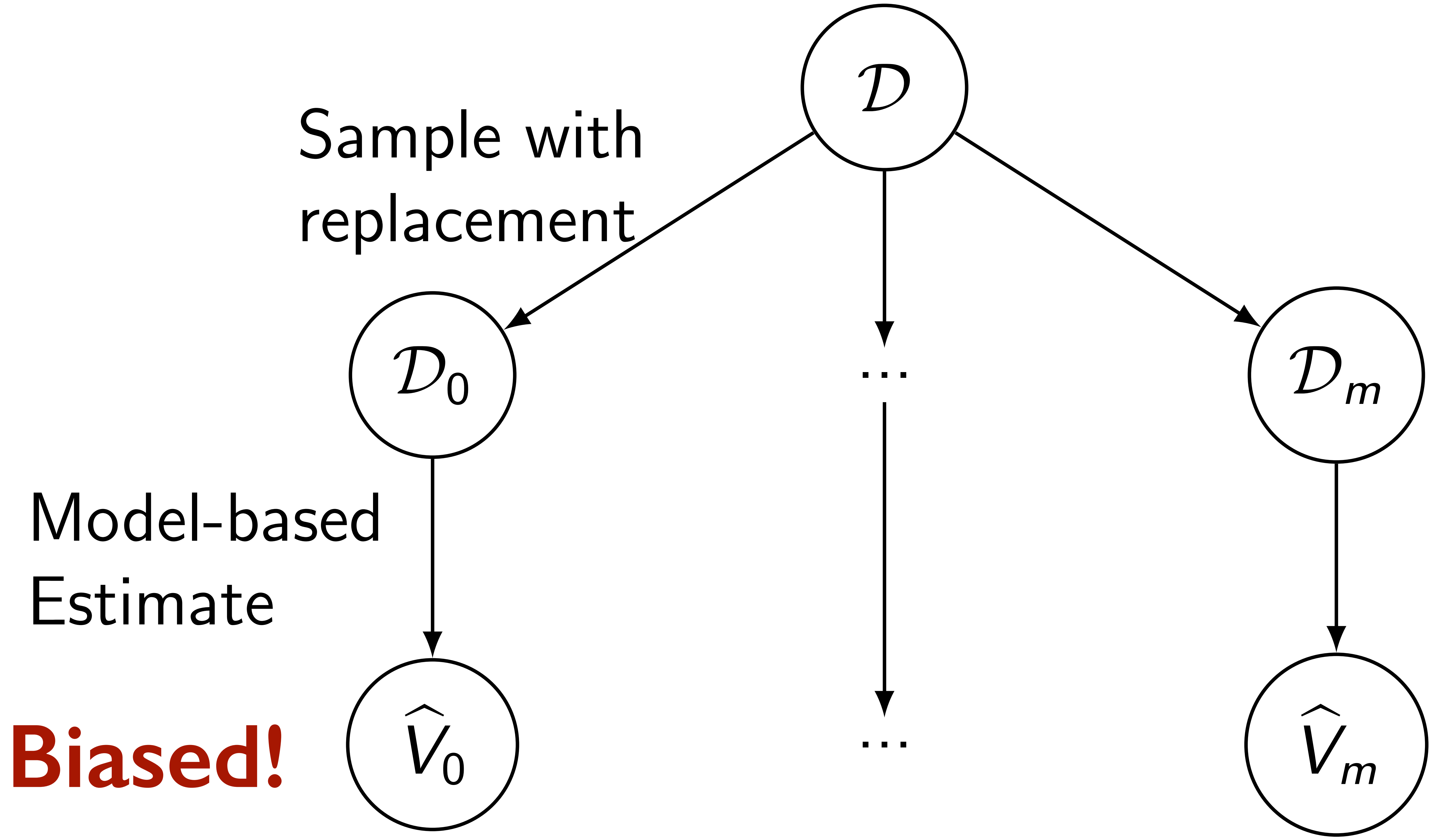
Bootstrap Confidence Intervals



Model-Based Bootstrap



Model-Based Bootstrap



Bad assumption #2:

"Biased models lead to biased estimators"

J.P. Hanna, P. Stone, and S. Niekum.
Bootstrapping with Models: Confidence Intervals for Off-Policy Evaluation.
International Conference on Autonomous Agents and Multiagent Systems (AAMAS), May 2017.

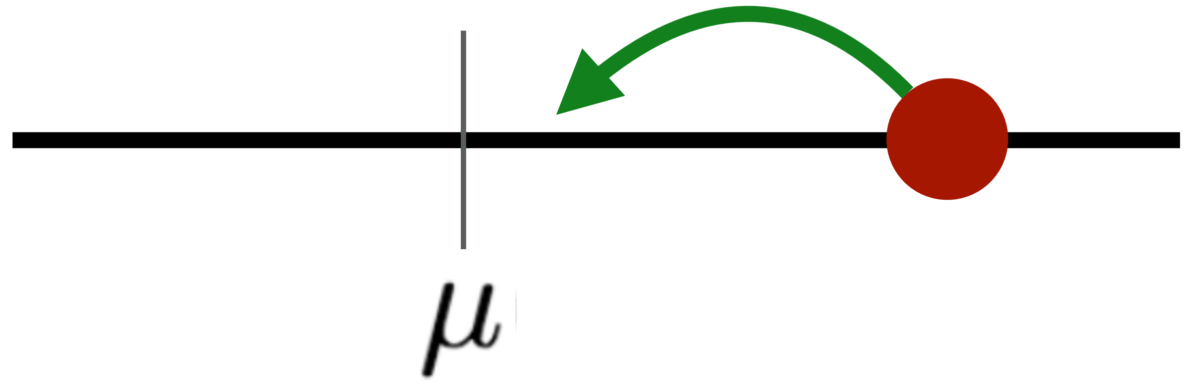
Doubly Robust Estimator

[Jiang and Li 2016; Thomas and Brunskill 2016]

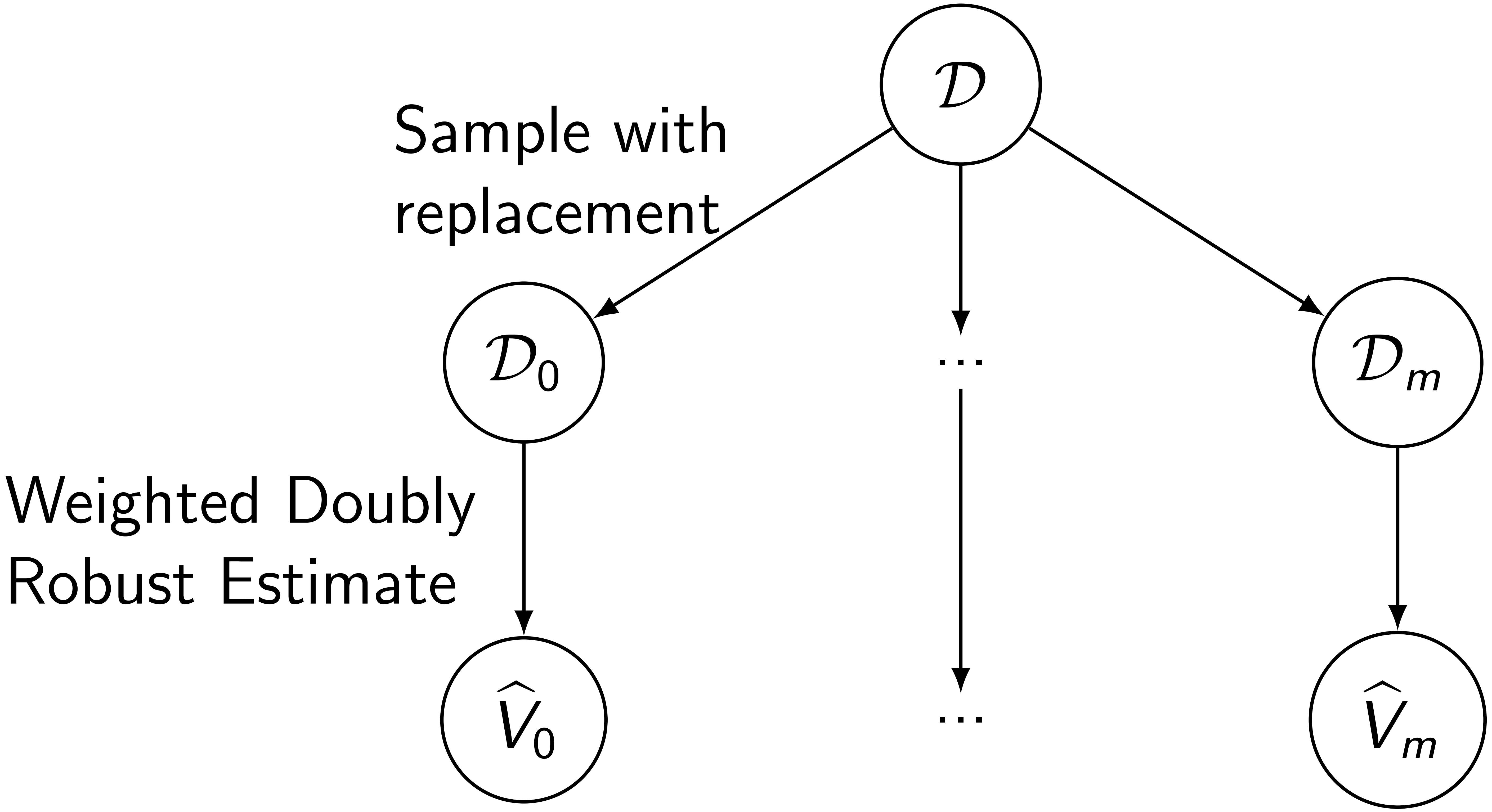
$$\mathrm{DR}(\mathcal{D}) := \underbrace{\mathrm{PDIS}(\mathcal{D})}_{\text{Unbiased estimator}} \underbrace{- \sum_{i=1}^{n} \sum_{t=0}^{L} w_t^i \hat{q}^{\pi_e}(S_t^i, A_t^i) - w_{t-1}^i \hat{v}^{\pi_e}(S_t^i)}_{\text{Zero in Expectation}}$$

Control variate ↑

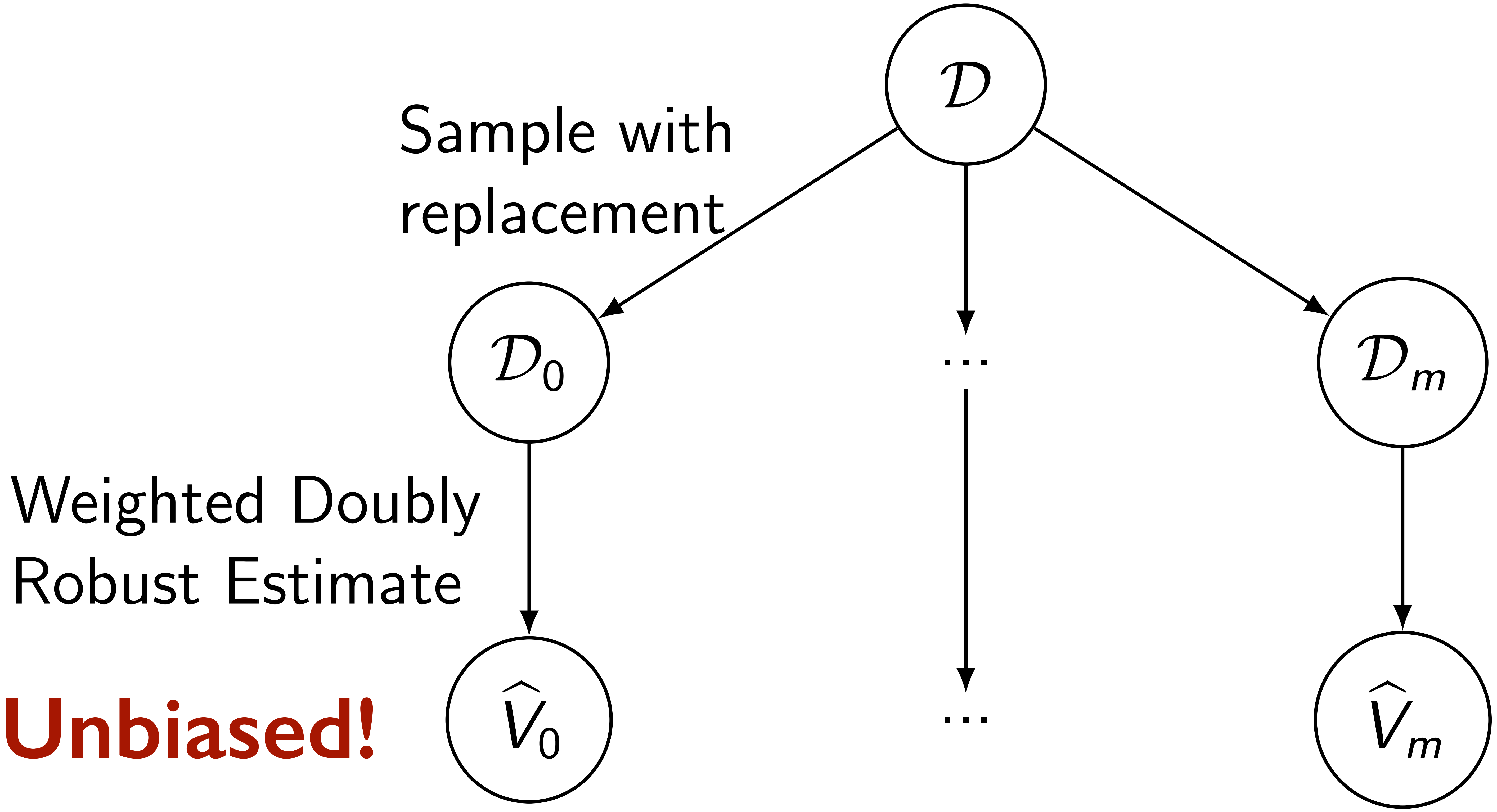
- $\hat{v}^{\pi}(S) := \mathbb{E}_{A\sim\pi, S'\sim\hat{P}(\cdot|S,A)}\left[r(S,A) + \hat{v}(S')\right]$
  - State value function.
- $\hat{q}^{\pi}(S,A) := r(S,A) + \mathbb{E}_{S'\sim P(\cdot|S,A)}\left[\hat{v}(S')\right]$
  - State-action value function.
- w_t is the importance weight of the first t time-steps.



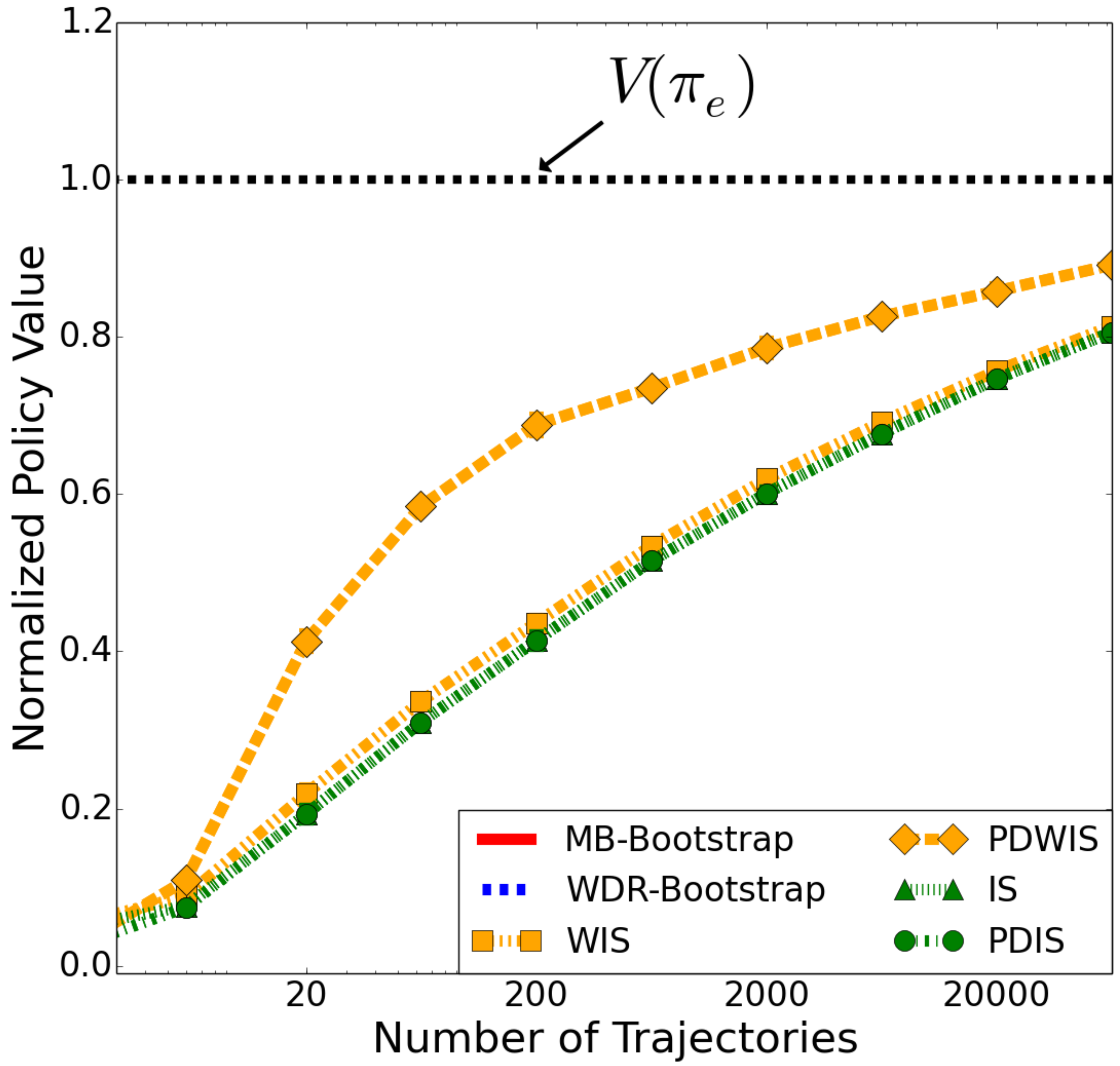
Weighted Doubly Robust Bootstrap



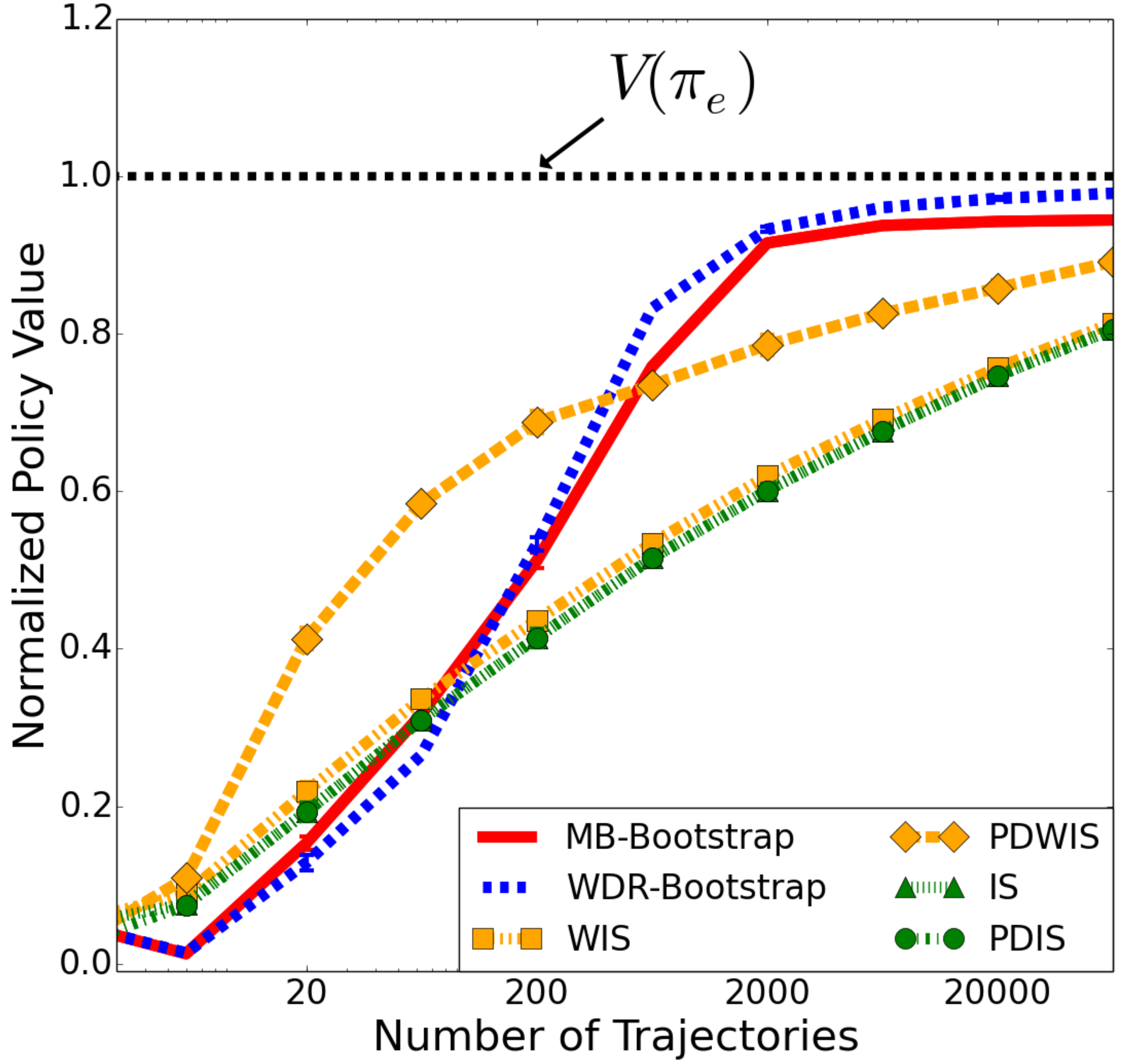
Weighted Doubly Robust Bootstrap



Mountain Car Results



Mountain Car Results



Similar ideas apply to safe policy improvement:

Loop:

1. Propose a policy (e.g. via an unsafe RL step)
2. Perform safe policy evaluation
3. Accept or reject

Putting it all together: Safe PI challenge problem

Hand coded walk (19.5 cm/s)

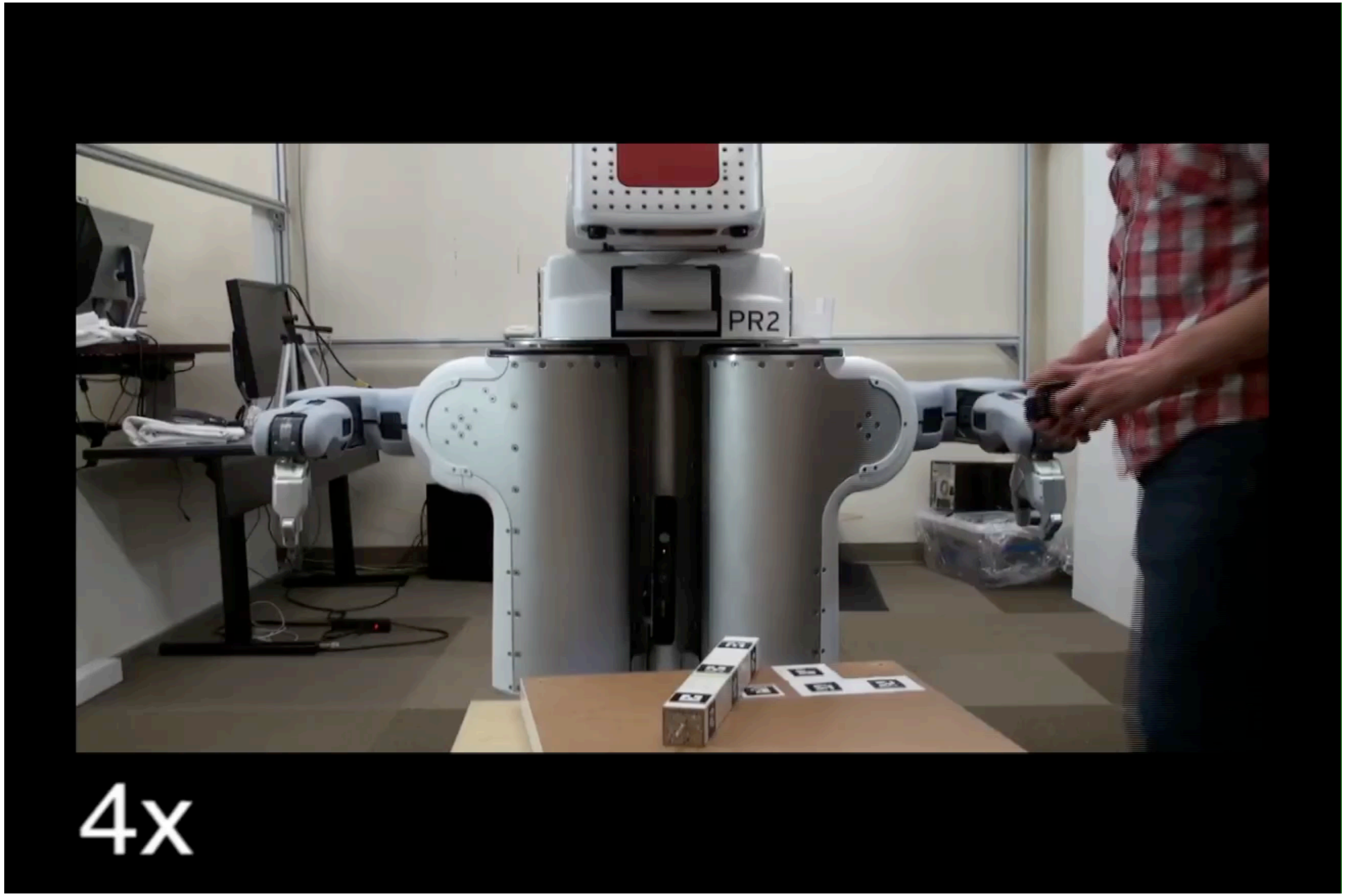
Best known walk (28 cm/s)

Without falling (more) during learning?

Part 1: Safe reinforcement learning

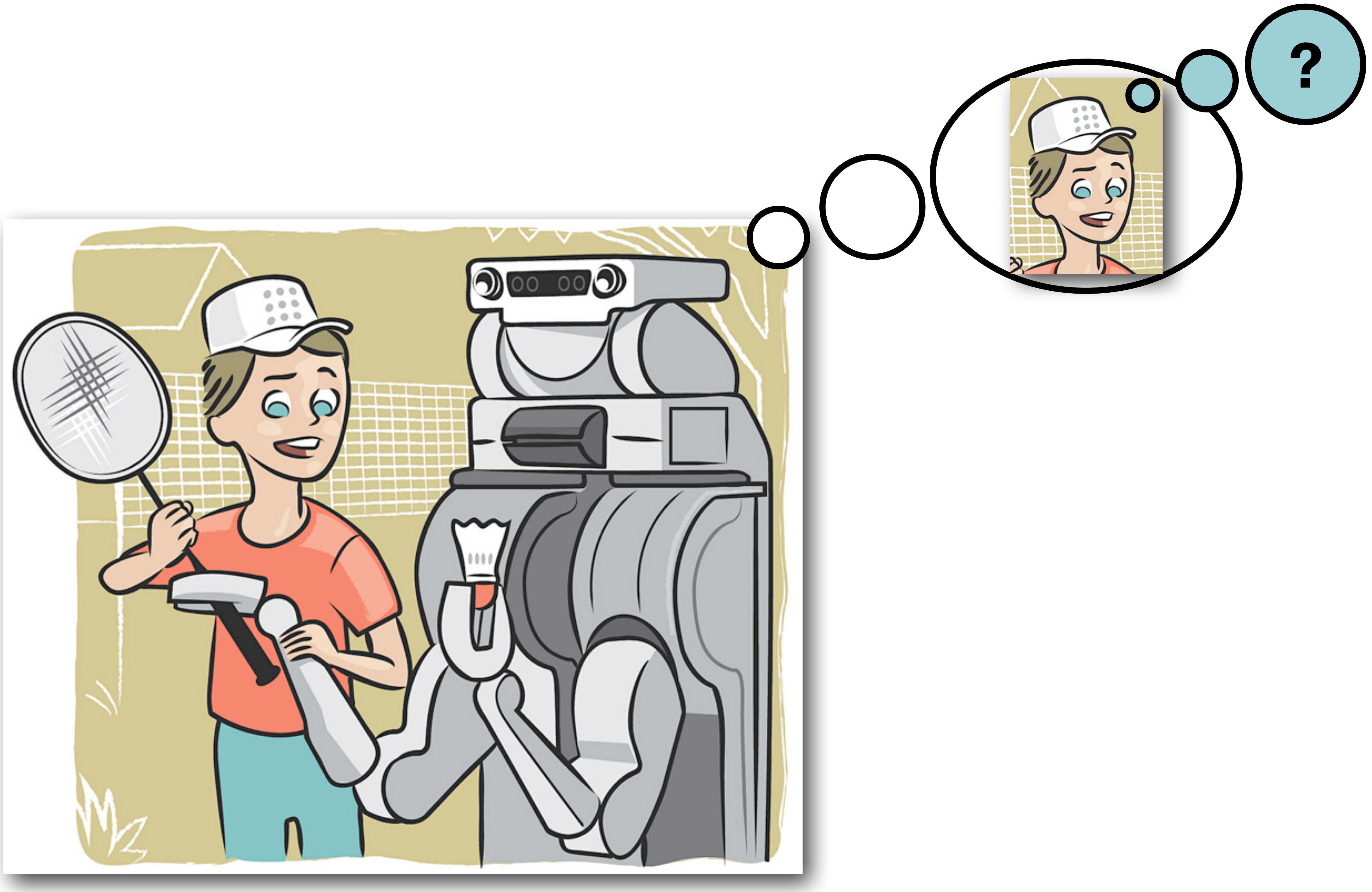
Part 2: Safe imitation learning

Imitation learning



Safe Imitation Learning:

Lower bound the **performance ratio** of the robot vs. human demonstrator with **high confidence**, *without knowing the ground-truth reward function.*



Inverse reinforcement learning: feature matching

(Abbeel and Ng 2004)

Policy value under linear reward function:
$$\begin{aligned} E_{s_0\sim D}[V^\pi(s_0)] &= E[\textstyle\sum_{t=0}^{\infty}\gamma^t R(s_t)|\pi] \\ &= E[\textstyle\sum_{t=0}^{\infty}\gamma^t w\cdot\phi(s_t)|\pi] \\ &= w\cdot E[\textstyle\sum_{t=0}^{\infty}\gamma^t\phi(s_t)|\pi] \end{aligned}$$

(Discounted) feature expectations:
$$\mu(\pi) = E[\textstyle\sum_{t=0}^{\infty}\gamma^t\phi(s_t)|\pi] \in \mathbb{R}^k.$$

Goal: find a reward function whose optimal policy matches expert's feature expectations

If expert's feature expectations are matched, then total return is also identical

Hoeffding-style bound (w.r.t. projection IRL algorithm)

(Abbeel and Ng 2004, Syed and Schapire 2008)

Theorem 2. *(Syed and Schapire 2008) To obtain a policy $\hat{\pi}$ such that with probability $(1-\delta)$*

$$\epsilon \geq |V^{\hat{\pi}}(R^*) - V^{\pi^*}(R^*)| \tag{26}$$

it suffices to have

$$m \geq \frac{2}{(\frac{\epsilon}{3}(1-\gamma))^2} \log \frac{2k}{\delta}. \tag{27}$$

Corollary 2. *Given a confidence level δ, and m demonstrations, with probability $(1-\delta)$ we have that $|V^{\pi^*}(R^*) - V^{\hat{\pi}}(R^*)| \leq \epsilon$, where*

$$\epsilon \leq \frac{3}{1-\gamma} \sqrt{\frac{2}{m} \log \frac{2k}{\delta}} \tag{28}$$

where k is the number of features and γ is the discount factor of the underlying MDP.

Bad assumption #3:

"Worst-case reasoning is the best we can do if we don't know the ground-truth reward function"

D.S. Brown and S. Niekum.
Efficient Probabilistic Performance Bounds for Inverse Reinforcement Learning.
AAAI Conference on Artificial Intelligence, February 2018.

D.S. Brown, Y. Cui, and S. Niekum.
Risk-Aware Active Inverse Reinforcement Learning.
Conference on Robot Learning (CoRL), October 2018.

Rethinking feature expectations

Problem 1: Hoeffding method bounds the features expectations, which in turn, bounds loss under a worst-case reward function, regardless of its likelihood given the demonstrations

Problem 2: Feature expectation methods cannot learn from state-action pairs that aren't part of a full trajectory

Bayesian Inverse Reinforcement Learning (BIRL)

[Ramachandran and Amir 2007]

- Use MCMC to sample from posterior:

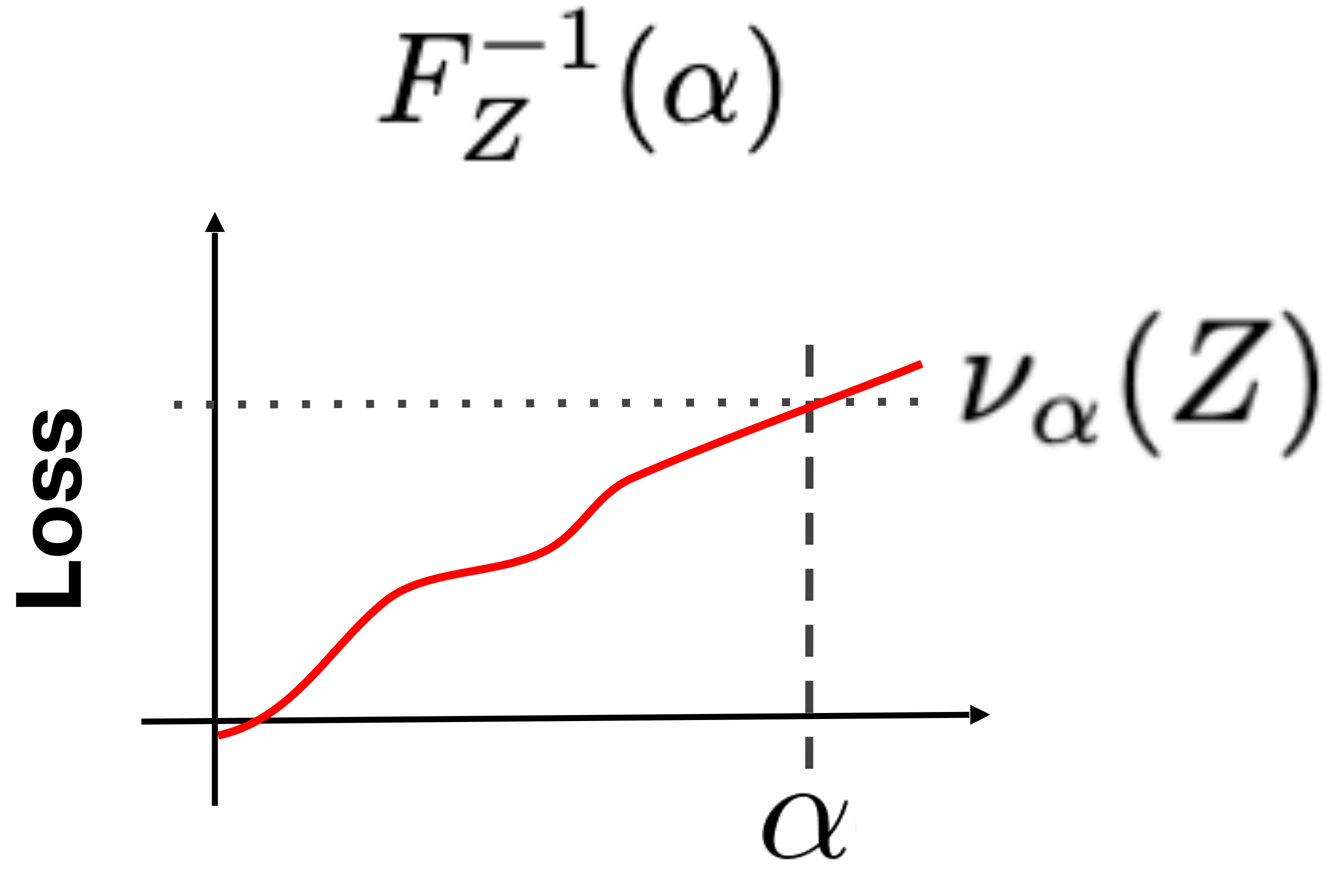
$$P(R|D) \propto P(D|R)P(R)$$

- Assume demonstrations follow softmax policy with temperature c:

$$P(D|R) = \prod_{(s,a)\in D} \frac{e^{cQ^*(s,a,R)}}{\sum_{b\in A} e^{cQ^*(s,b,R)}}$$

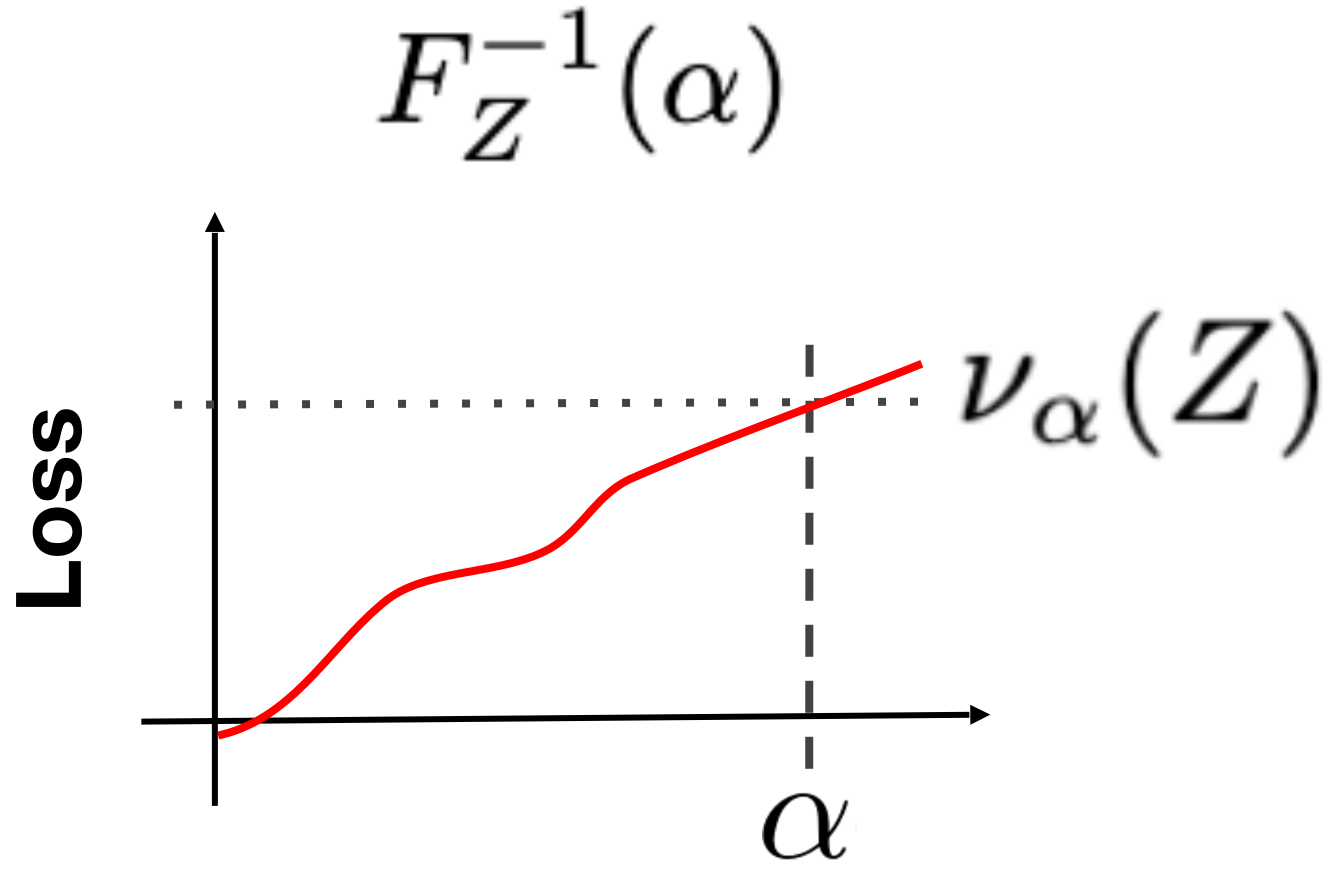
Value at risk

$$\nu_\alpha(Z) = F_Z^{-1}(\alpha) = \inf\{z : F_Z(z) \geq \alpha\}$$



Value at risk

$$\nu_\alpha(Z) = F_Z^{-1}(\alpha) = \inf\{z : F_Z(z) \geq \alpha\}$$



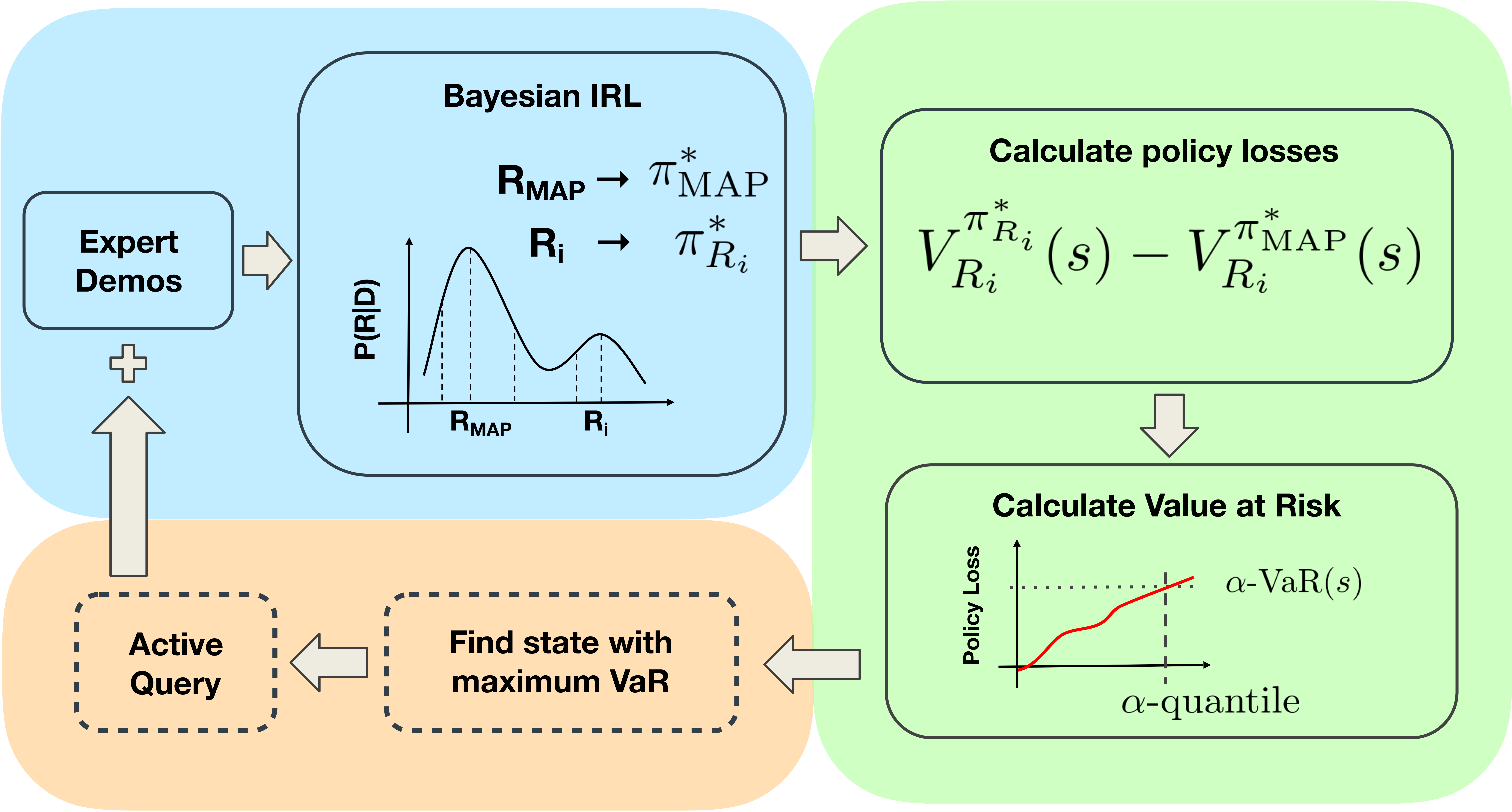
+

Single-sided confidence bound

"With probability $1-\delta$, no more than $1-\alpha\%$ of the outcomes will be worse than X"

Goal: Solve for X and check if it is below acceptable risk level

(Active) Safe IRL Pipeline



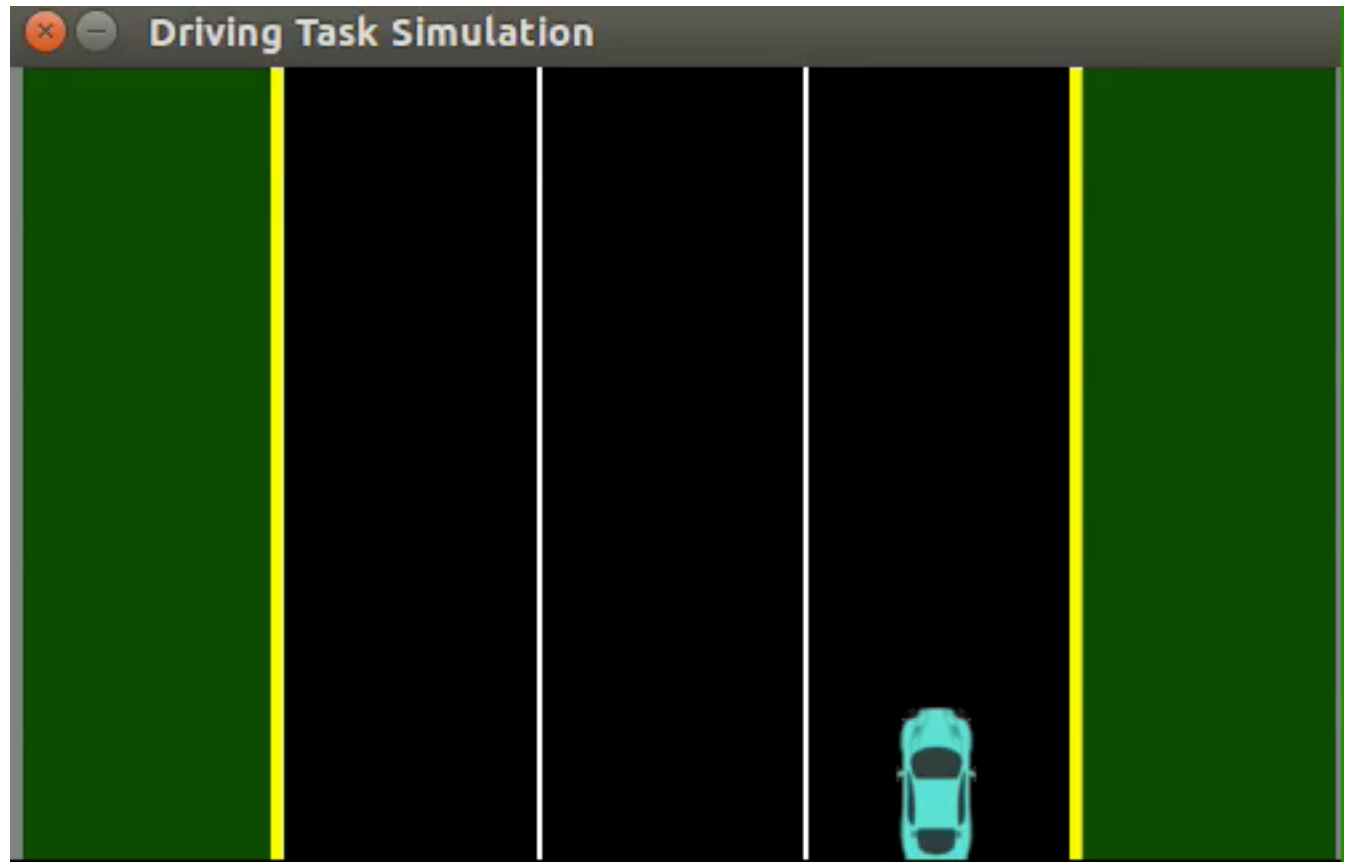
Results: efficiency (no active learning)

	Number of demonstrations					Average Accuracy
	1	5	9	⋯	23,146	
0.95-VaR EVD Bound	**0.9372**	**0.2532**	**0.1328**		-	0.98
0.99-VaR EVD Bound	1.1428	0.2937	0.1535		-	1.0
EVD Bound (Syed and Schapire 2008)	142.59	63.77	47.53		0.9372	1.0

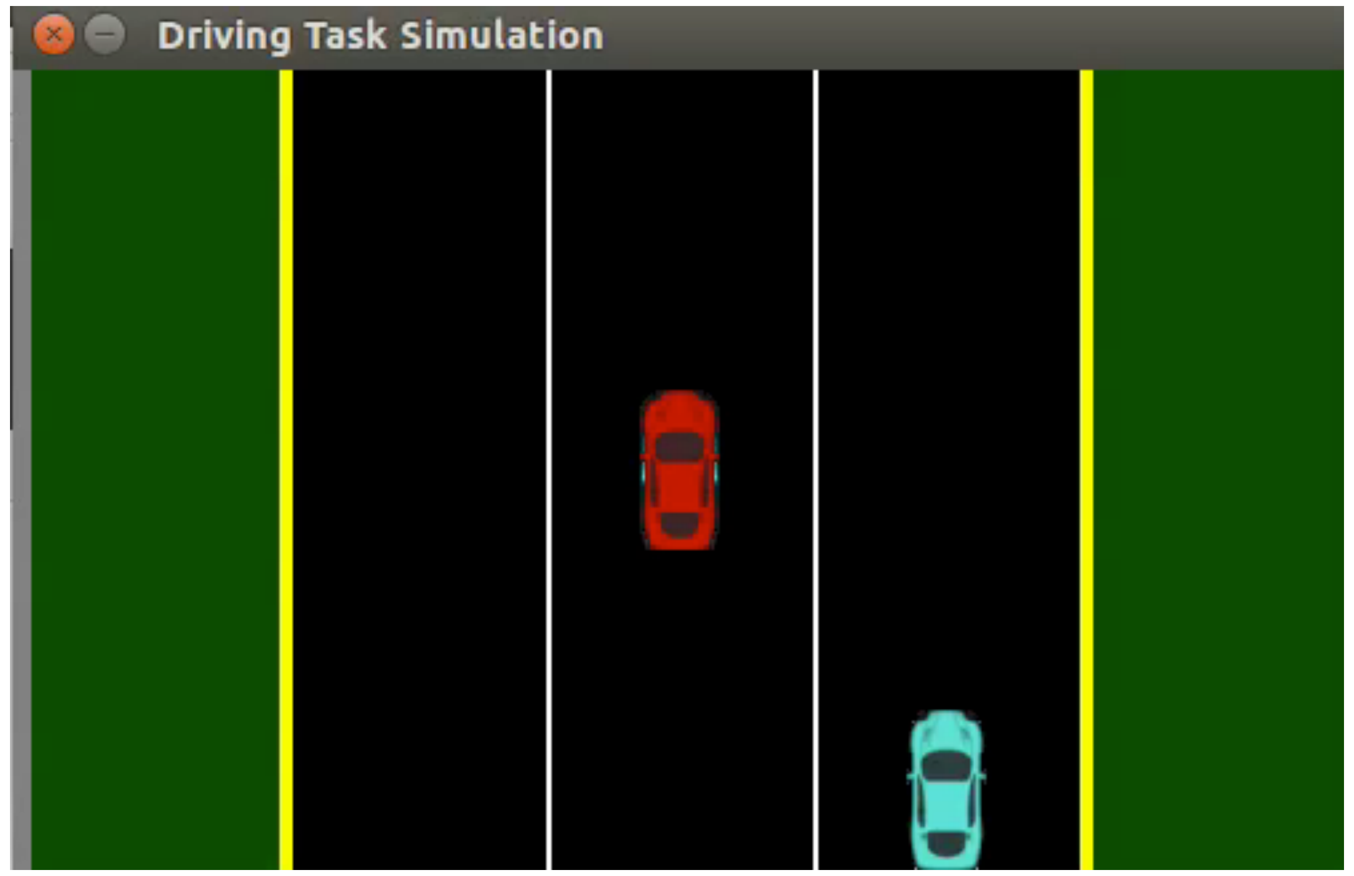
Table 1: Comparison of 95% confidence α-VaR bounds with a 95% confidence Hoeffding-style bound (Syed and Schapire 2008). Both bounds use the Projection algorithm (Abbeel and Ng 2004) to obtain the evaluation policy. Results are averaged over 200 random navigation tasks.

Four orders of magnitude more data efficient!

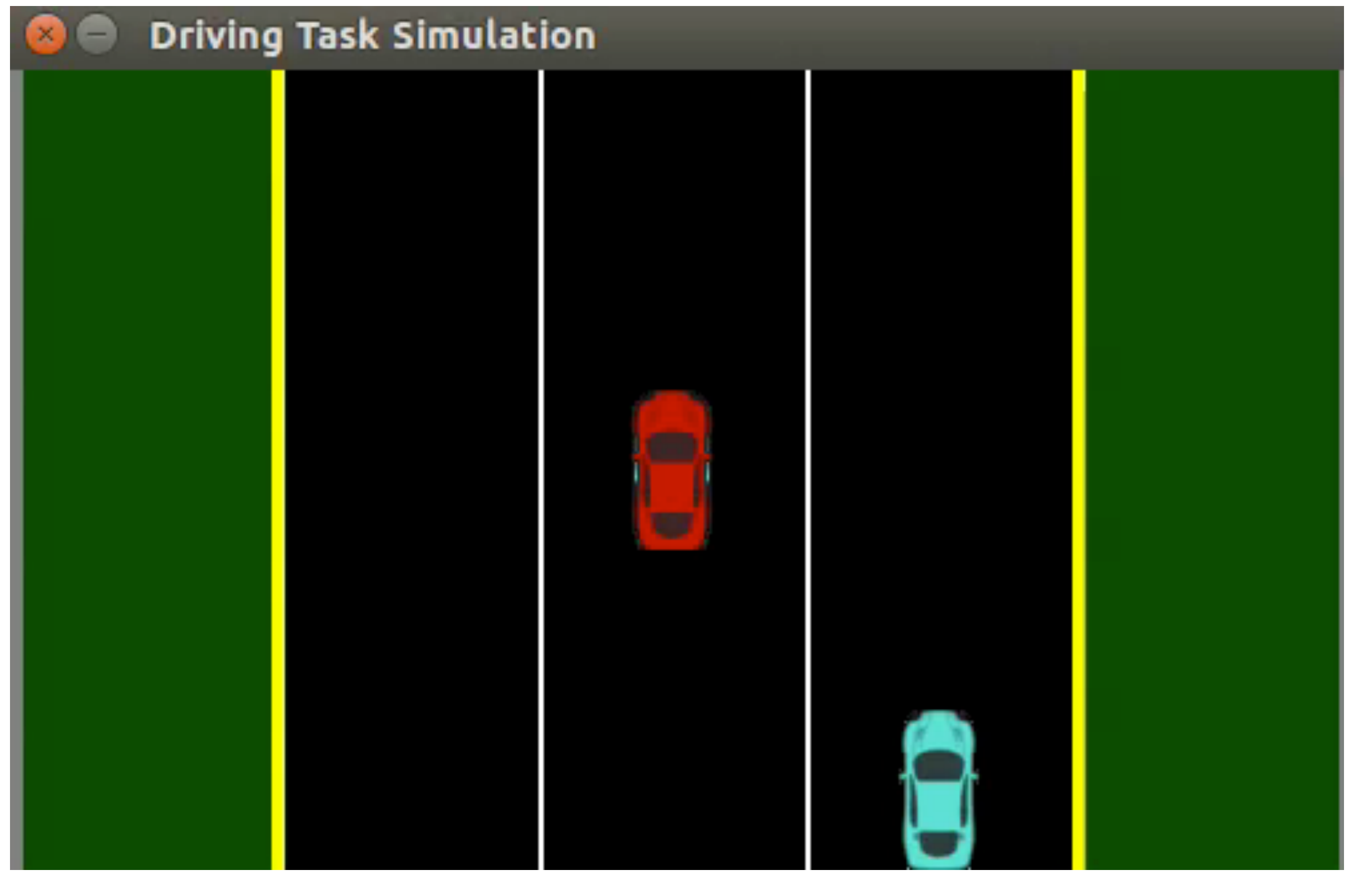
Risk-sensitive preferences



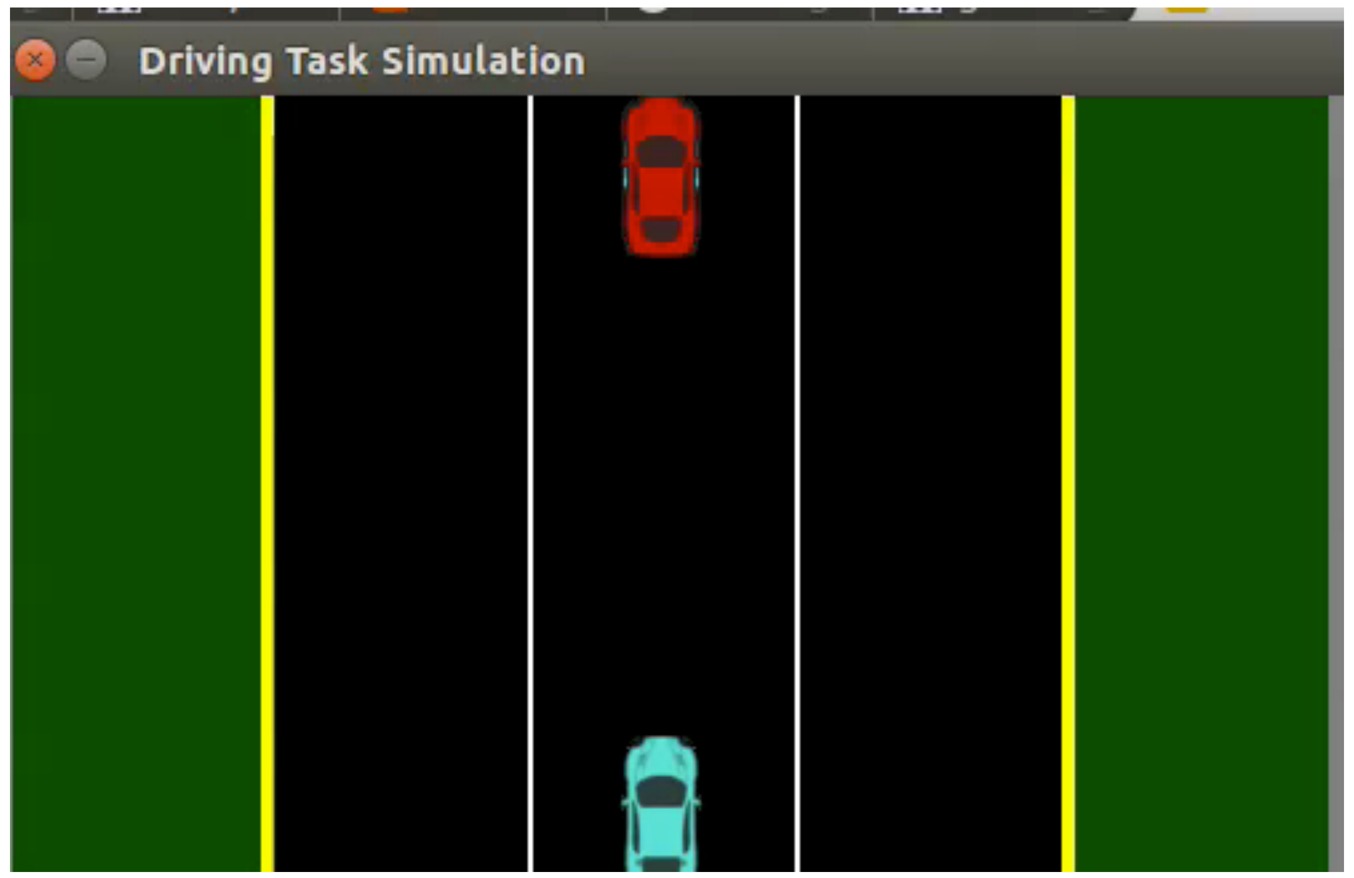
Demonstration: avoids cars, no lane pref



Avoids cars, but prefers right lane

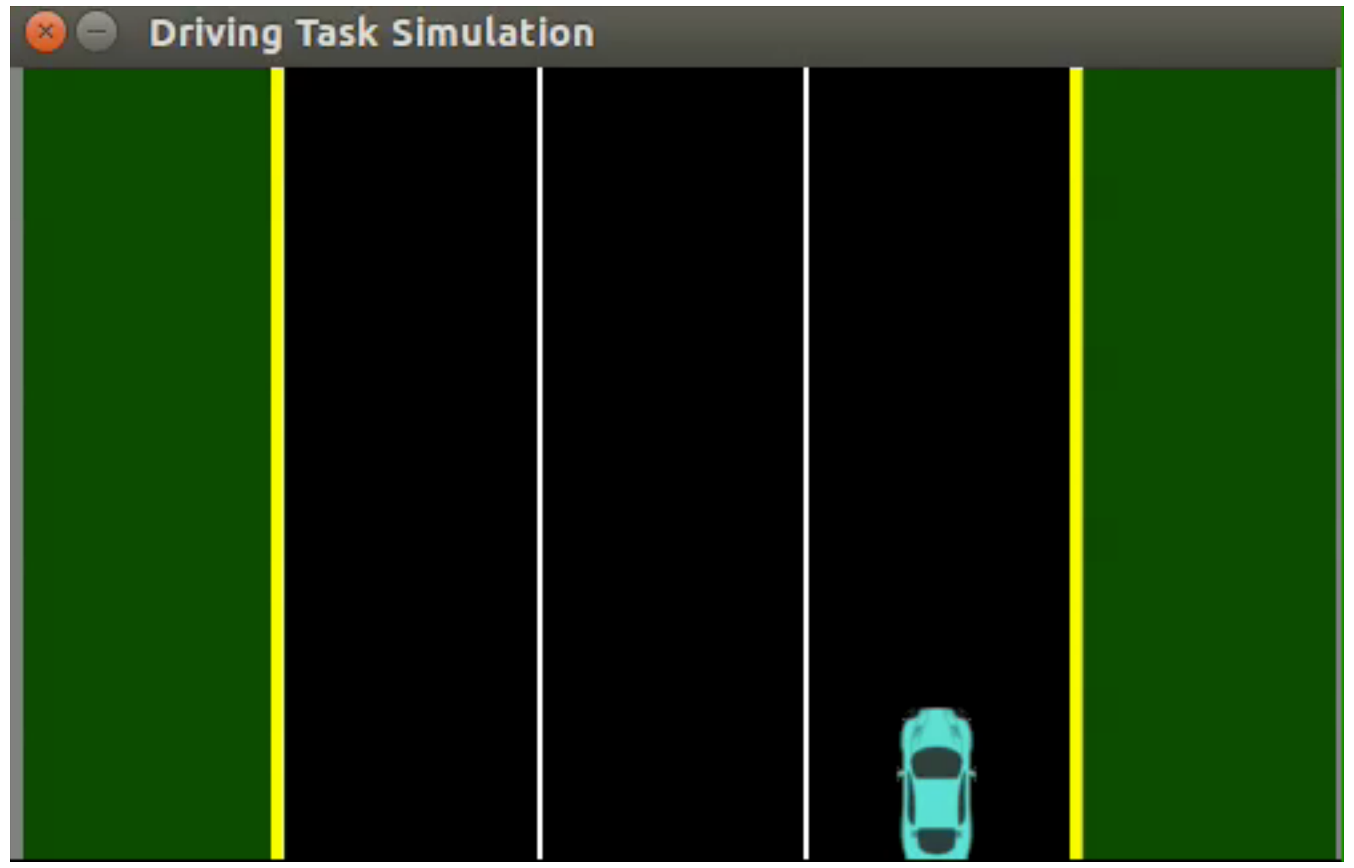


Stays on road, but ignores other cars



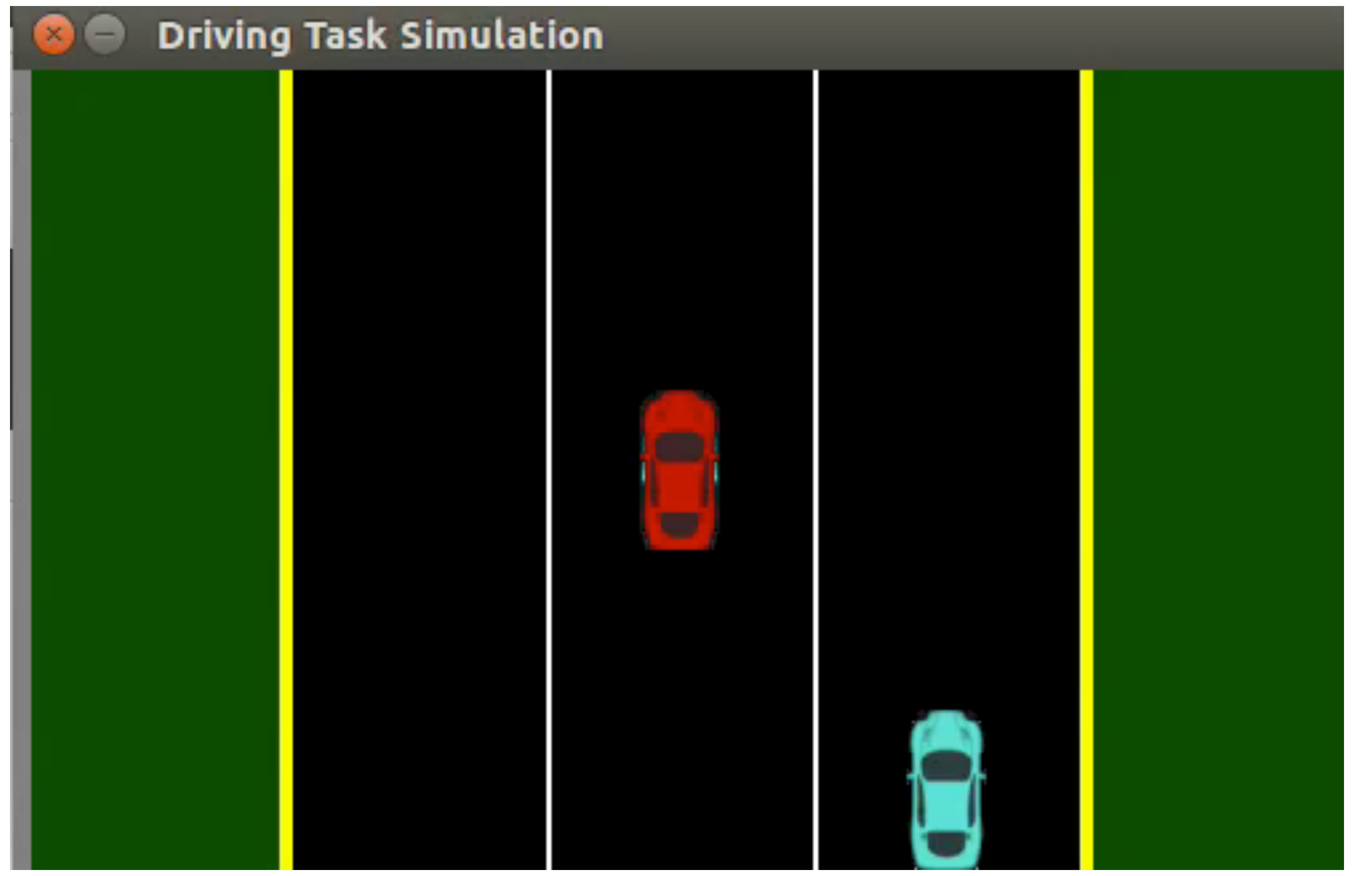
Seeks collisions

Risk-sensitive preferences (feature count-based)



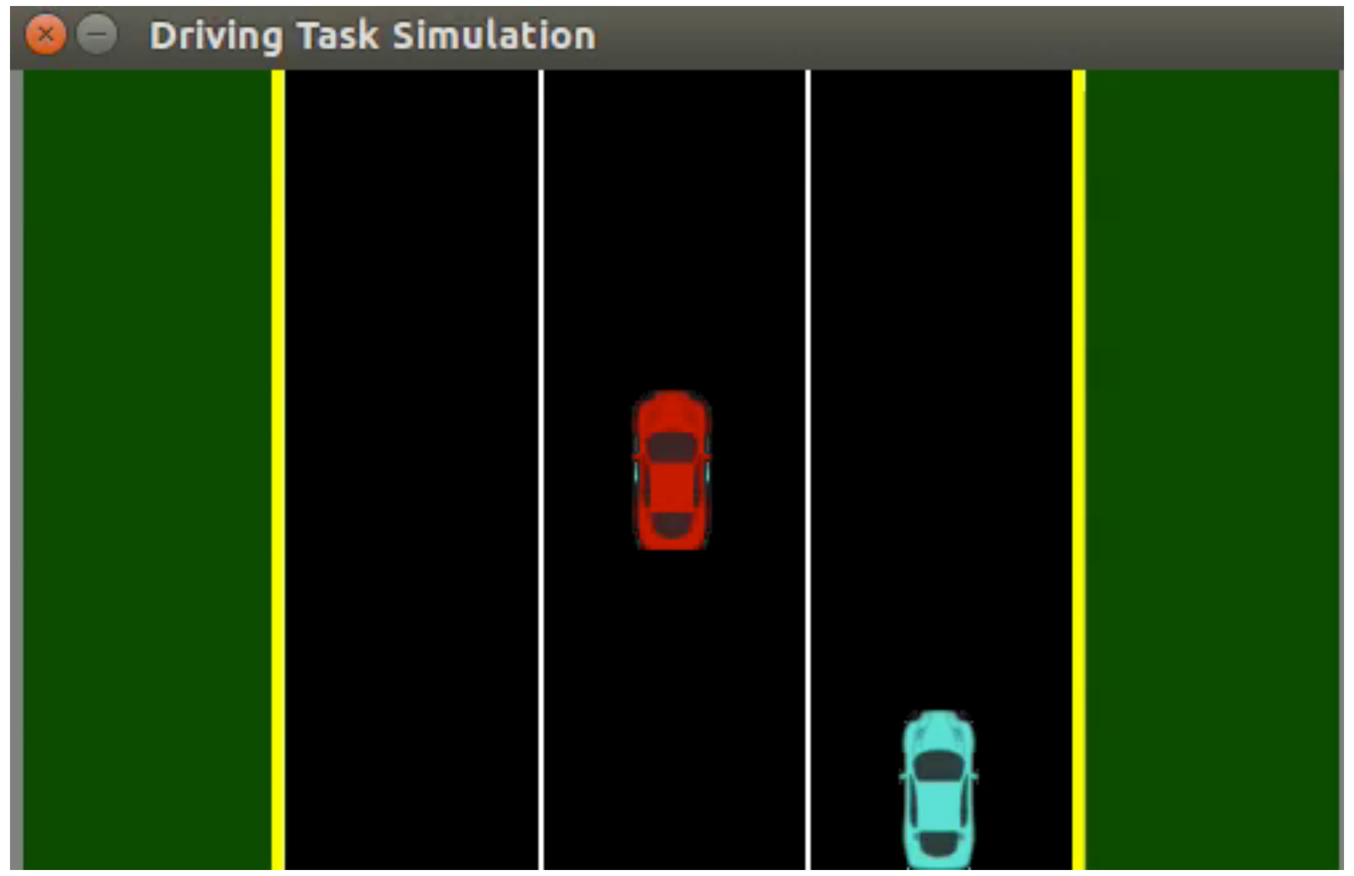
Demonstration: avoids cars, no lane pref

3



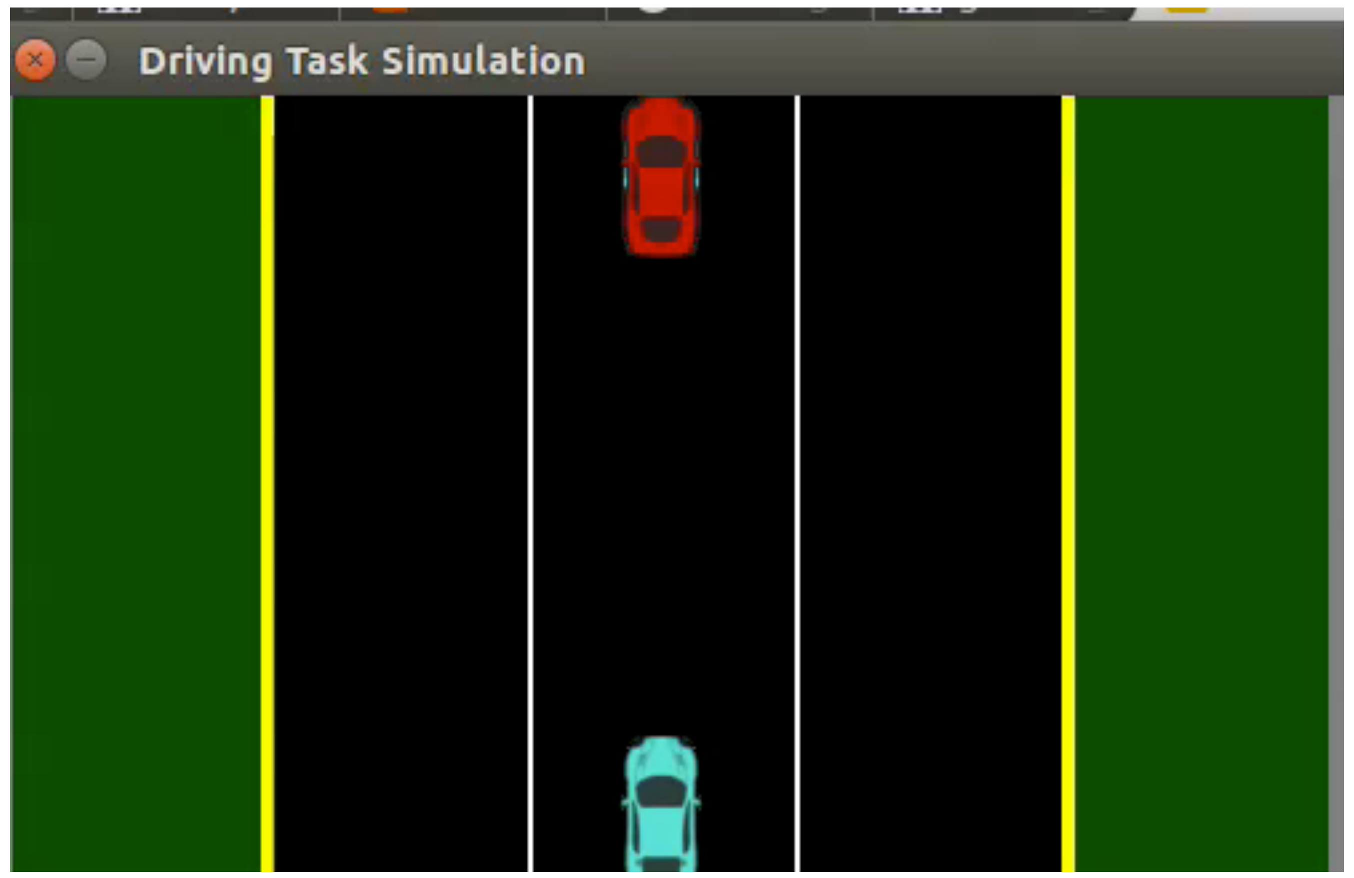
Avoids cars, but prefers right lane

1



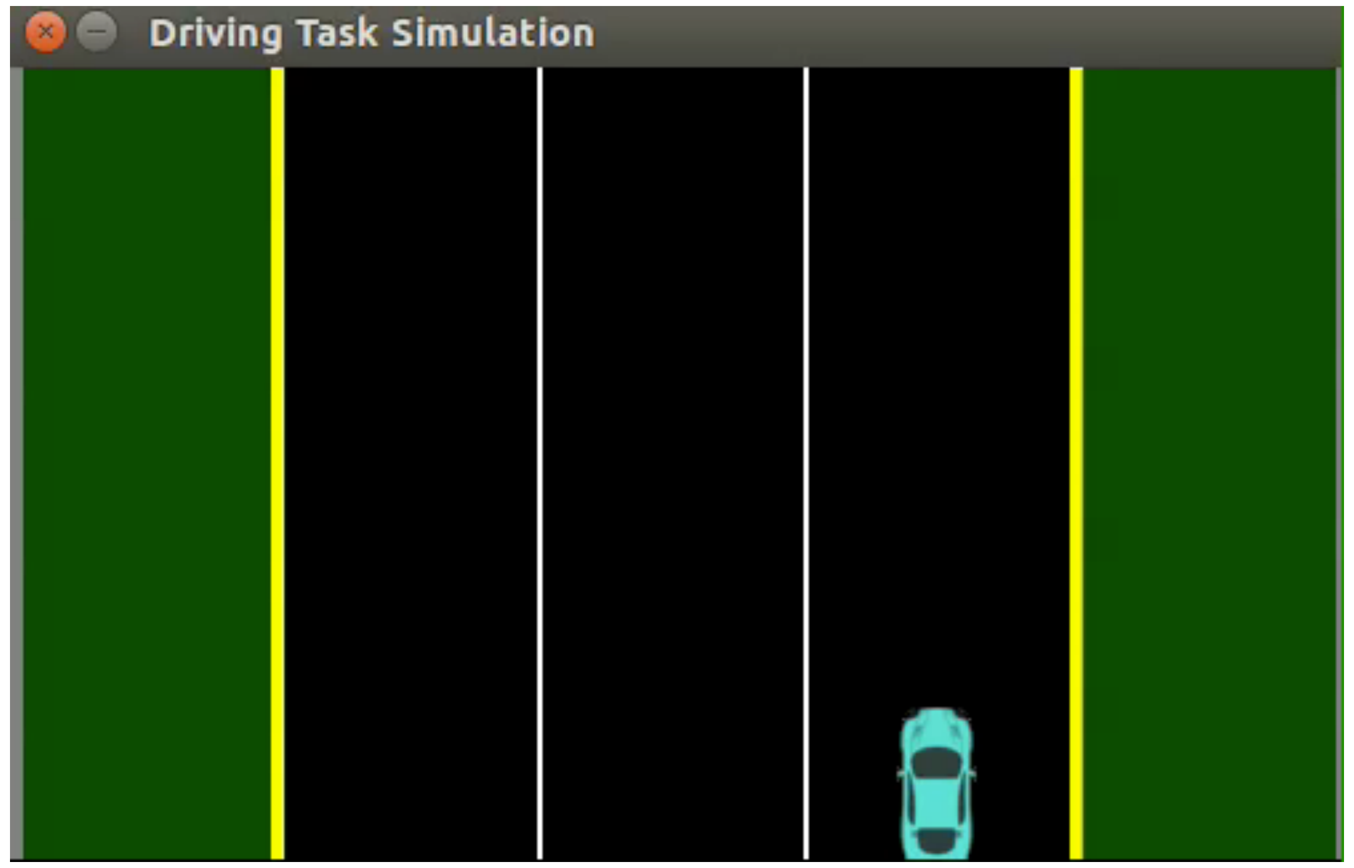
Stays on road, but ignores other cars

2



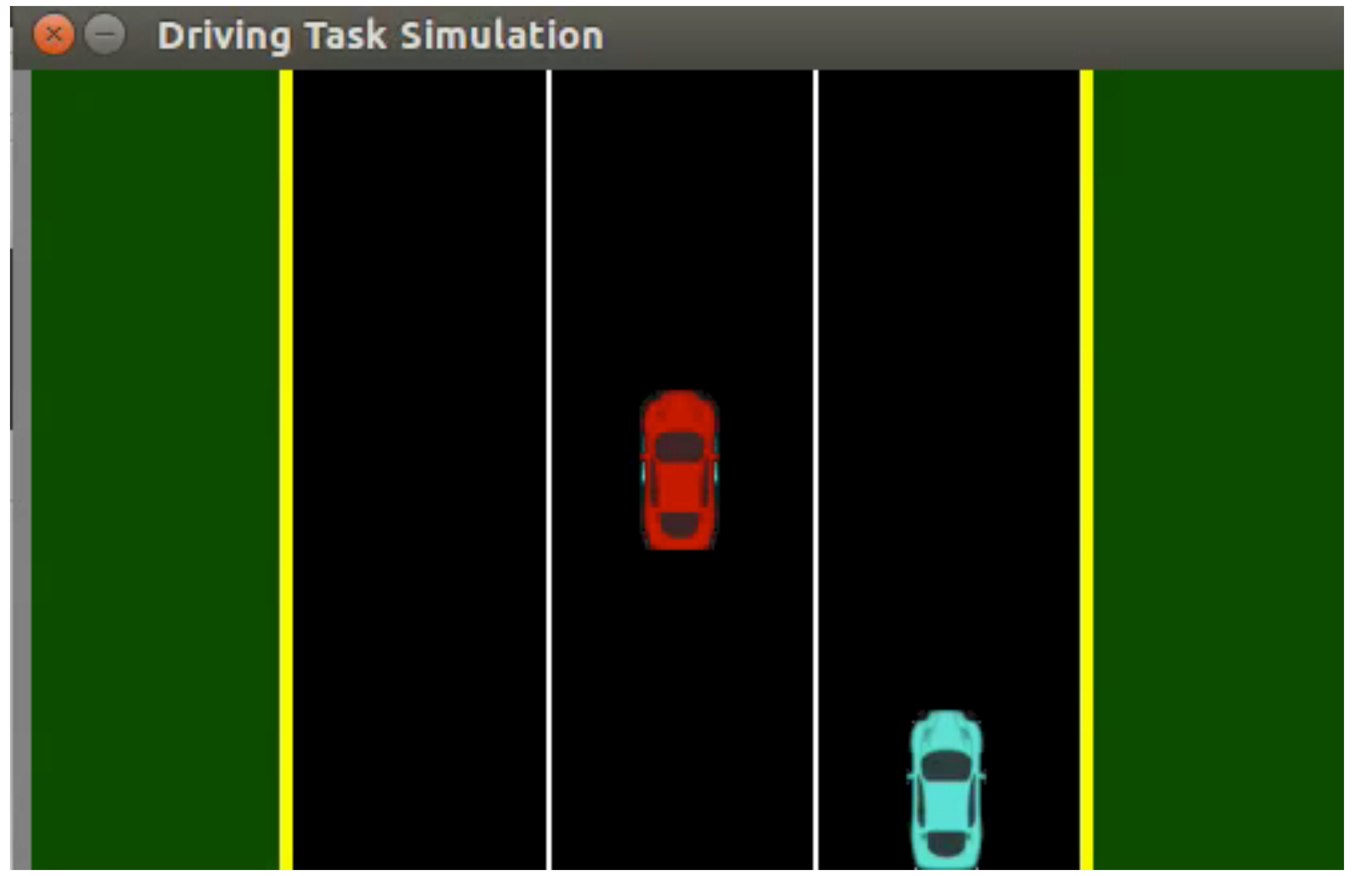
Seeks collisions

Risk-sensitive preferences (our approach)



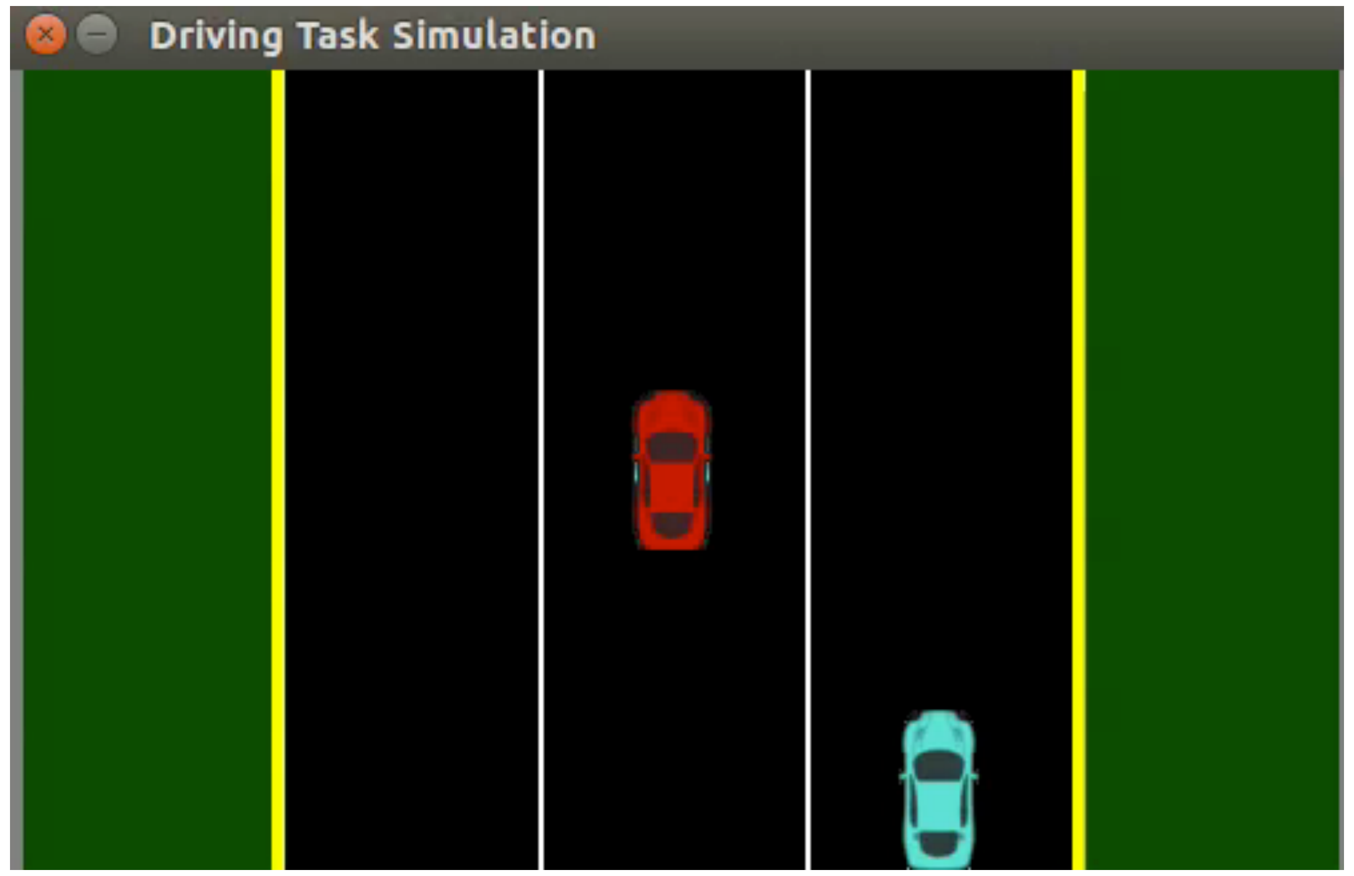
Demonstration: avoids cars, no lane pref

1



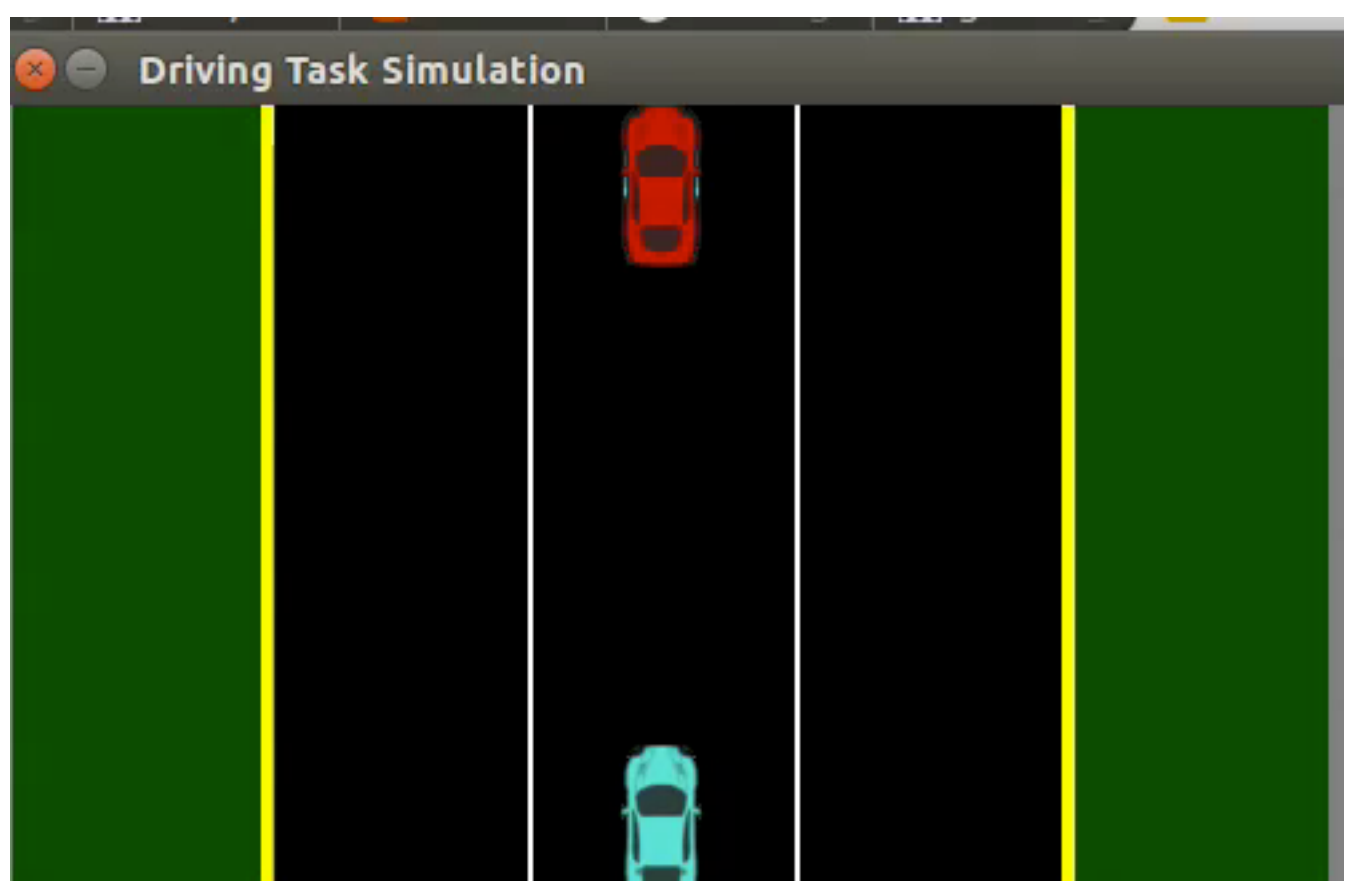
Avoids cars, but prefers right lane

2



Stays on road, but ignores other cars

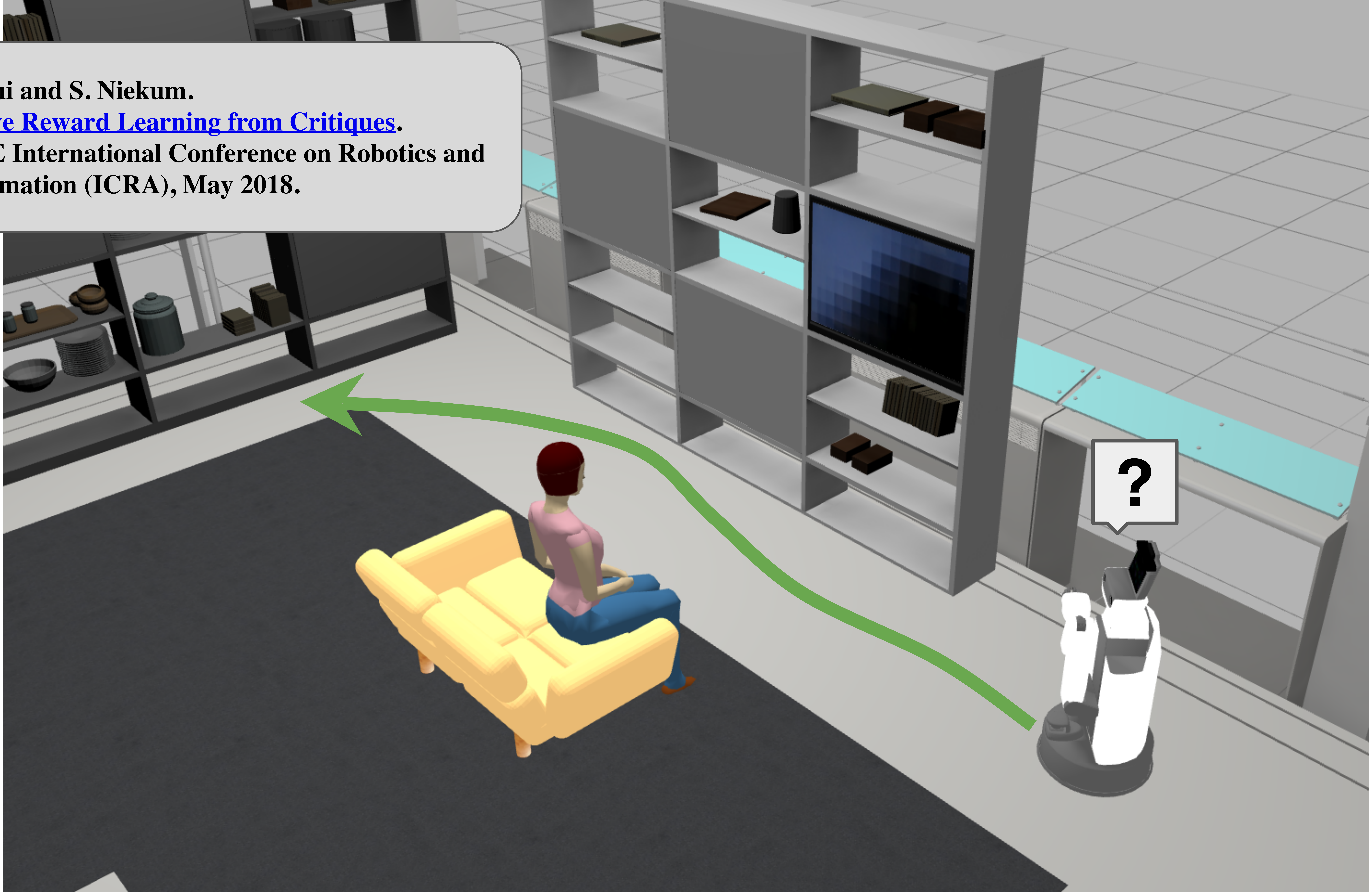
3



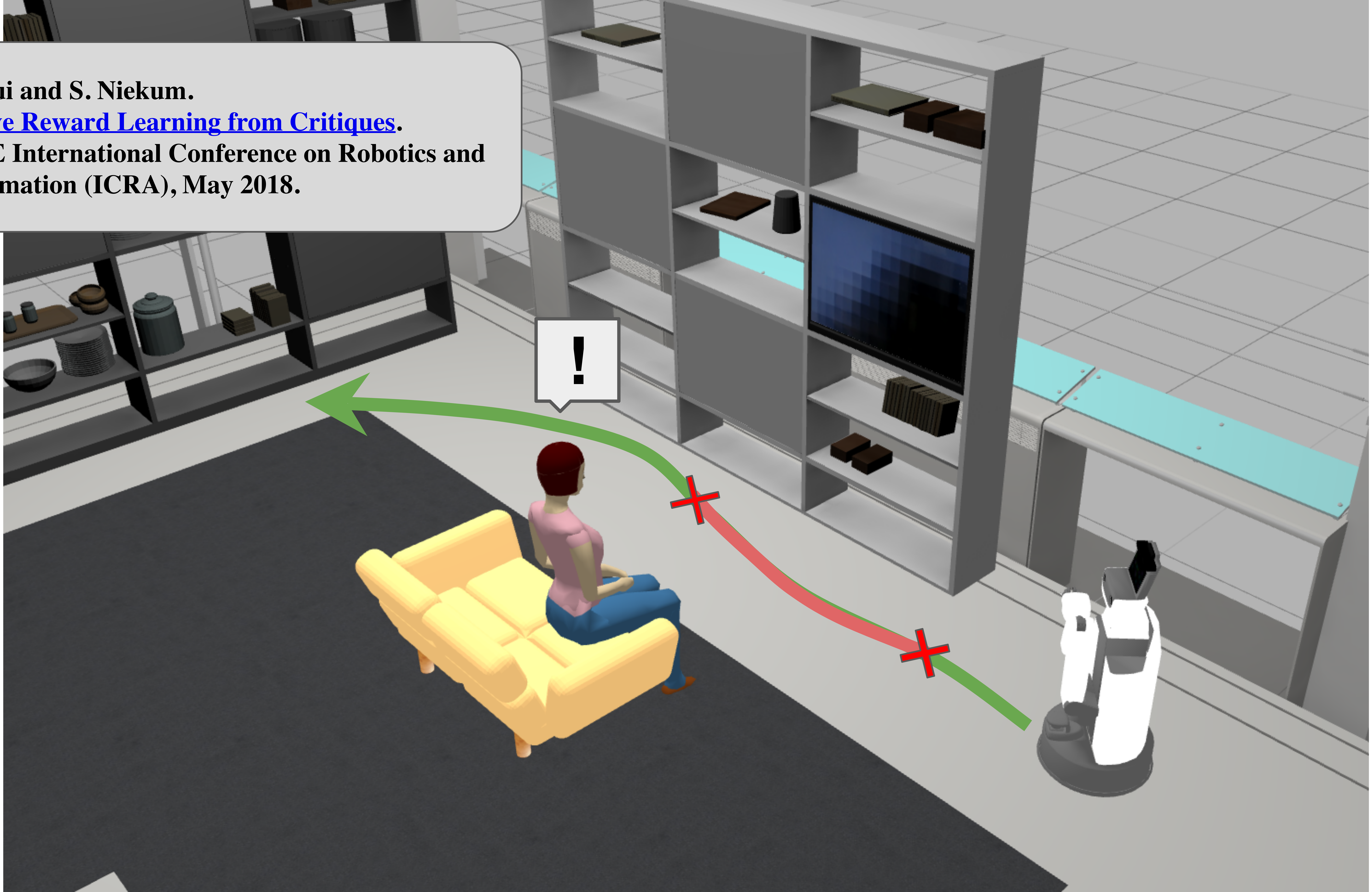
Seeks collisions

Risk-sensitive policy search

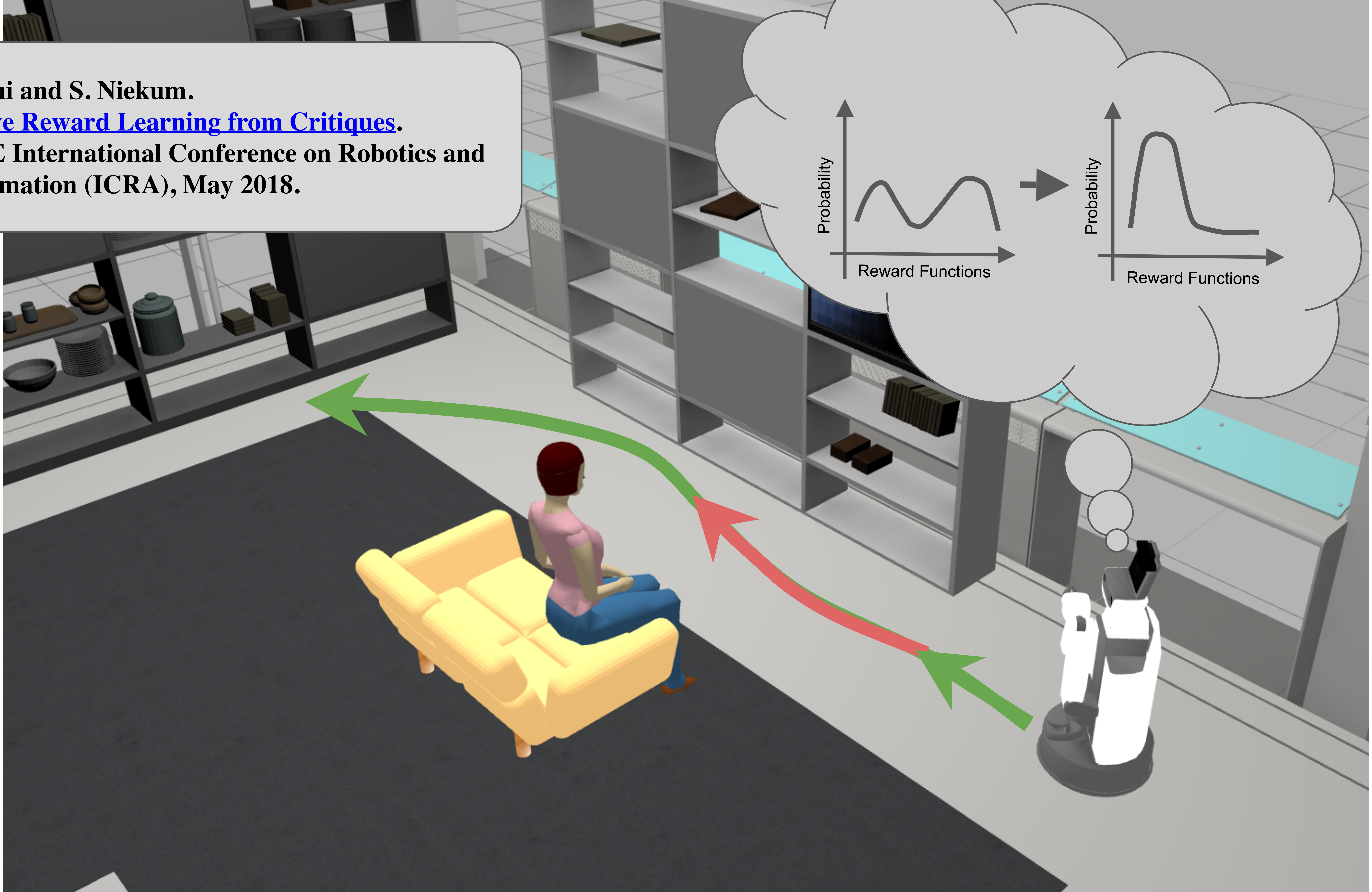
Y. Cui and S. Niekum.
[Active Reward Learning from Critiques]().
IEEE International Conference on Robotics and Automation (ICRA), May 2018.



Y. Cui and S. Niekum.
Active Reward Learning from Critiques.
IEEE International Conference on Robotics and Automation (ICRA), May 2018.

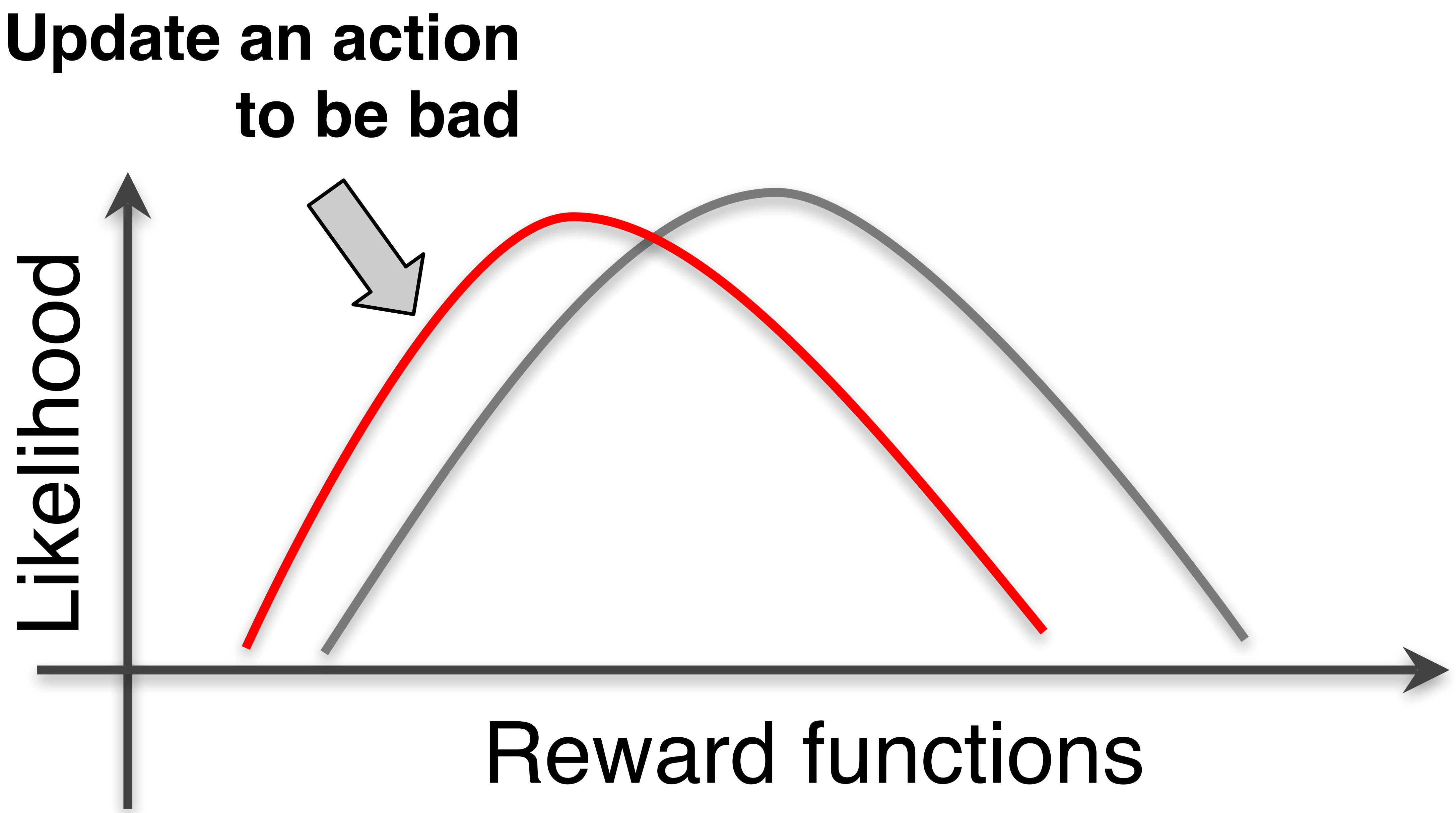


Y. Cui and S. Niekum.
Active Reward Learning from Critiques.
IEEE International Conference on Robotics and Automation (ICRA), May 2018.

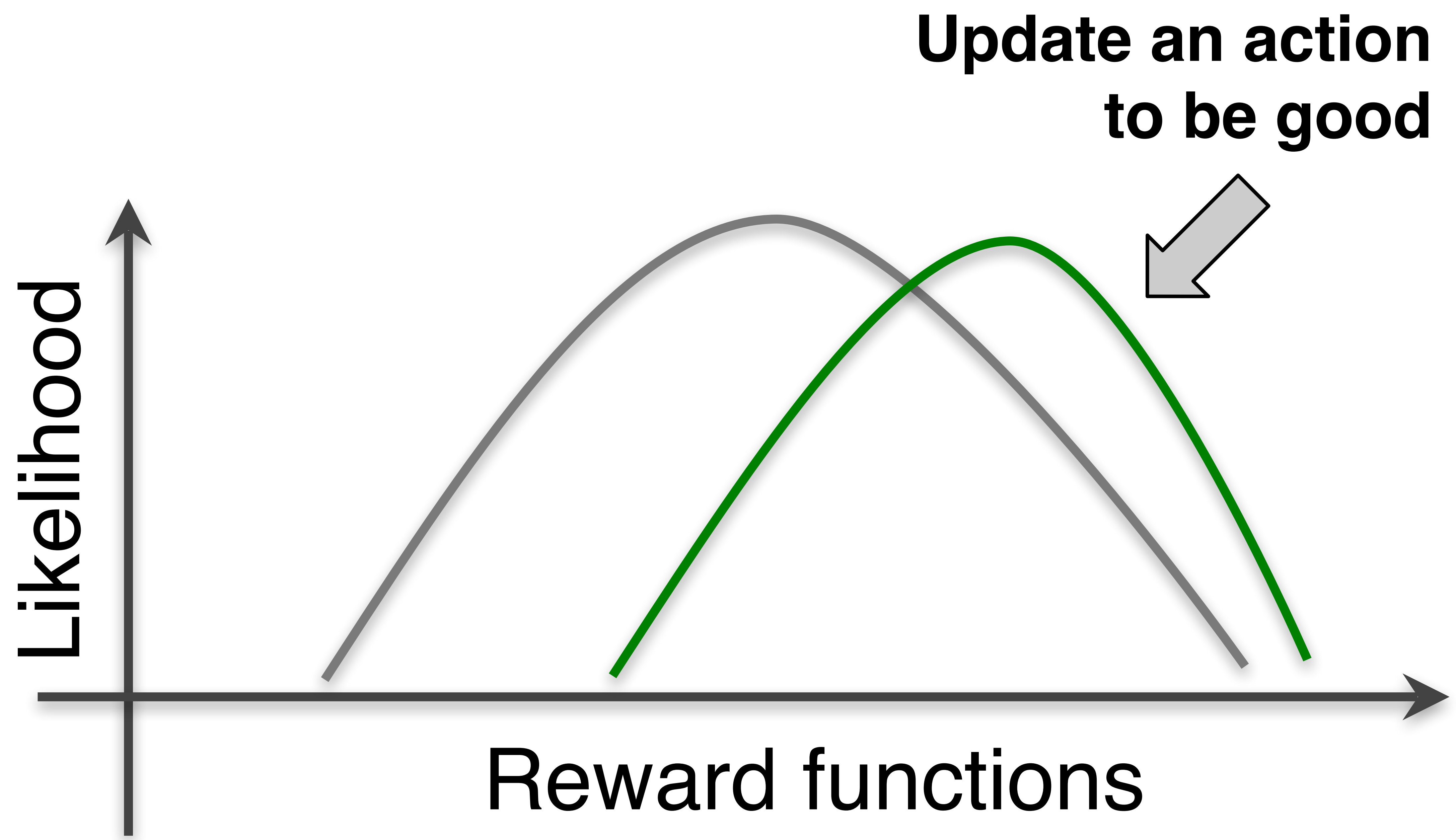


Information Gain Estimation from Reward Function Distribution

$$Pr(a_i \notin O(s_i) \mid R) = 1 - \frac{1}{Z_i} e^{\alpha Q(s_i, a_i, R)}$$



$$Pr(a_i \in O(s_i) \mid R) = \frac{1}{Z_i} e^{\alpha Q(s_i, a_i, R)}$$



Information Gain Estimation from Reward Function Distribution

$$Pr(a_i \notin O(s_i) \mid R) = 1 - \frac{1}{Z_i} e^{\alpha Q(s_i, a_i, R)}$$

$$Pr(a_i \in O(s_i) \mid R) = \frac{1}{Z_i} e^{\alpha Q(s_i, a_i, R)}$$

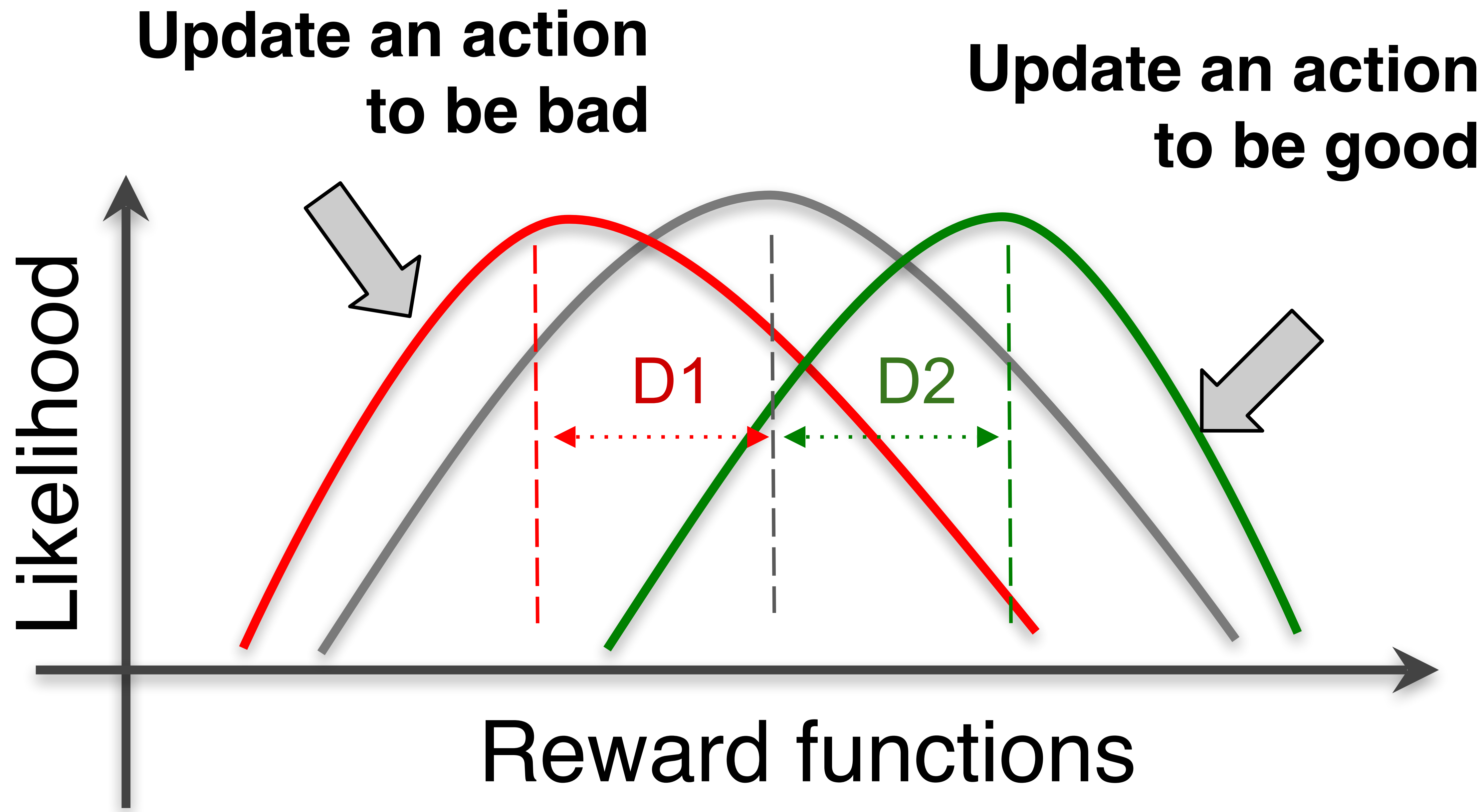
- Set of optimal actions at a state:

$$O(s) = \arg\max_{a \in A} Q^{\pi}(s, a)$$

- Distance Measure:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

- Expected Information Gain:



$$G^{+}(s_i, a_i) = G(D^{+} \cup (s_i, a_i) \mid Be(R)) = Pr(a_i \in O(s_i) \mid Be(R)) D(Be'(R)||Be(R))$$

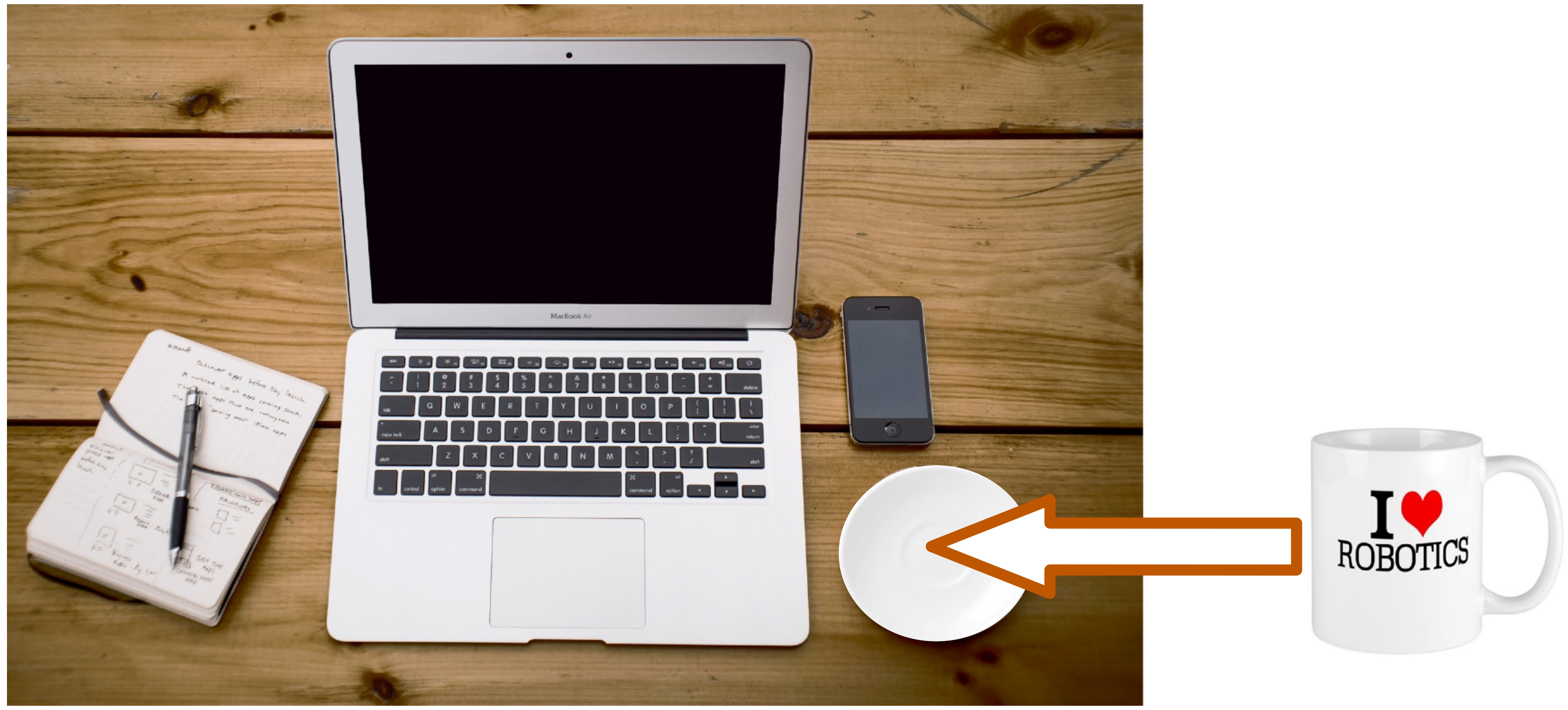
$$G^{-}(s_i, a_i) = G(D^{-} \cup (s_i, a_i) \mid Be(R)) = Pr(a_i \notin O(s_i) \mid Be(R)) D(Be'(R)||Be(R))$$

Bad assumption #4:

"Demonstration data should be treated as I.I.D."

D.S. Brown and S. Niekum.
Machine Teaching for Inverse Reinforcement Learning: Algorithms and Applications.
AAAI Conference on Artificial Intelligence, February 2019.

Informative demonstrations



Less informative

More informative

Machine teaching

In general:

$$\min_{D} \quad \text{TeachingCost}(D)$$
$$s.t. \quad \text{TeachingRisk}(\hat{\theta}) \leq \epsilon$$
$$\hat{\theta} = \text{MachineLearning}(D)$$

For inverse RL:

$$\min_{\mathcal{D}} \quad \text{TeachingCost}(\mathcal{D})$$
$$s.t. \quad \text{Loss}(\mathbf{w}^*, \hat{\mathbf{w}}) \leq \epsilon$$
$$\hat{\pi} = \text{RL}(\hat{\mathbf{w}})$$
$$\hat{\mathbf{w}} = \text{IRL}(\mathcal{D})$$

where:

$$\text{Loss}(\mathbf{w}^*, \hat{\mathbf{w}}) = \mathbf{w}^{*T}(\mu_{\pi^*} - \mu_{\hat{\pi}})$$

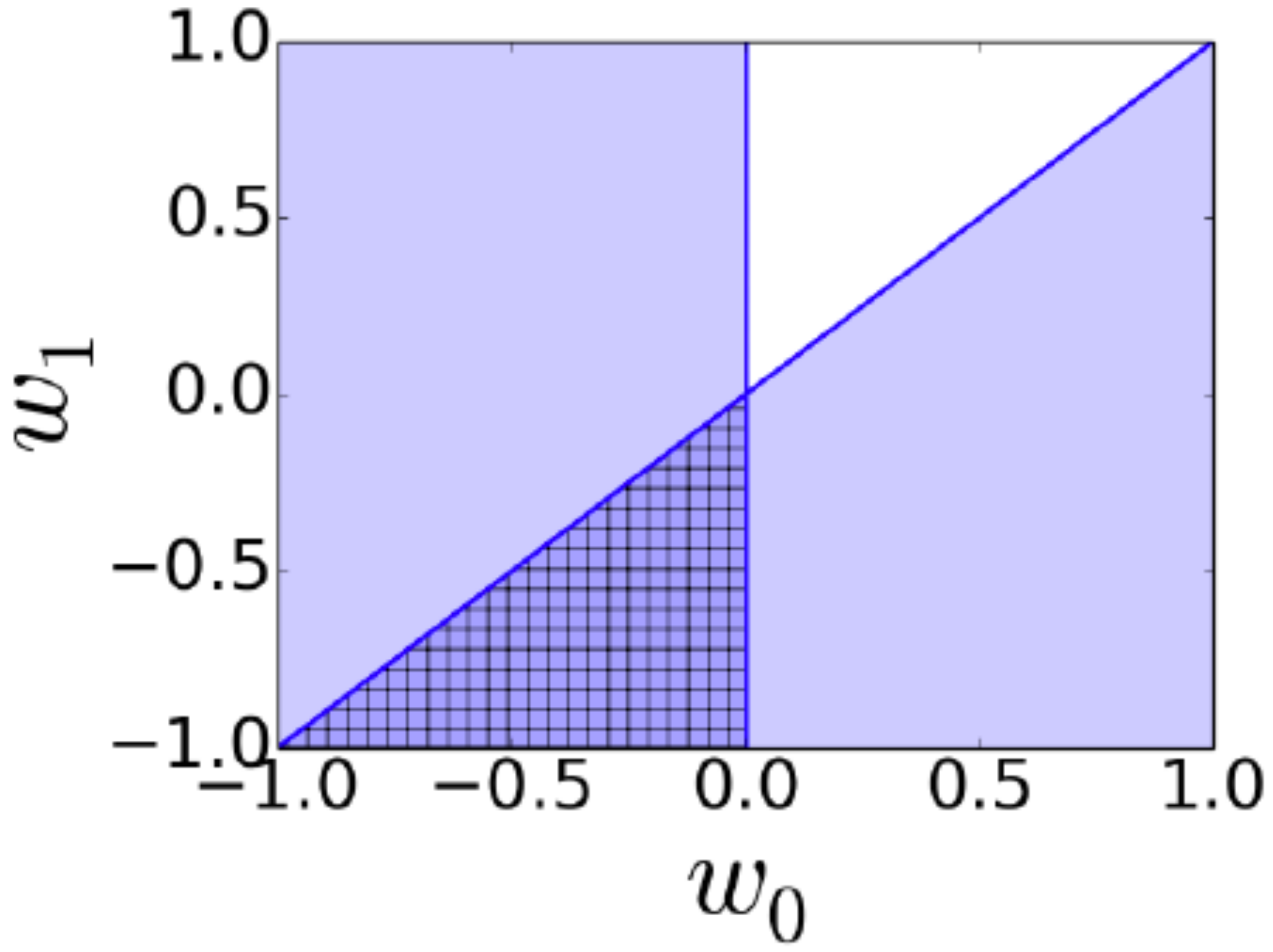
$$\text{TeachingCost}(\mathcal{D}) = |\mathcal{D}|$$

Behavioral Equivalence Classes (BEC)

$$\text{BEC}(\pi) =$$
$$\{\mathbf{w} \in \mathbb{R}^k \mid \pi \text{ is optimal under } R(s) = \mathbf{w}^T \phi(s)\}.$$

Theorem 1. *(Ng and Russell 2000) Given an MDP, BEC(π) is given by the following intersection of half-spaces:*

$$\mathbf{w}^T(\mu_\pi^{(s,a)} - \mu_\pi^{(s,b)}) \geq 0,$$
$$\forall a \in \arg\max_{a' \in \mathcal{A}} Q^*(s, a'), b \in \mathcal{A}, s \in \mathcal{S}$$

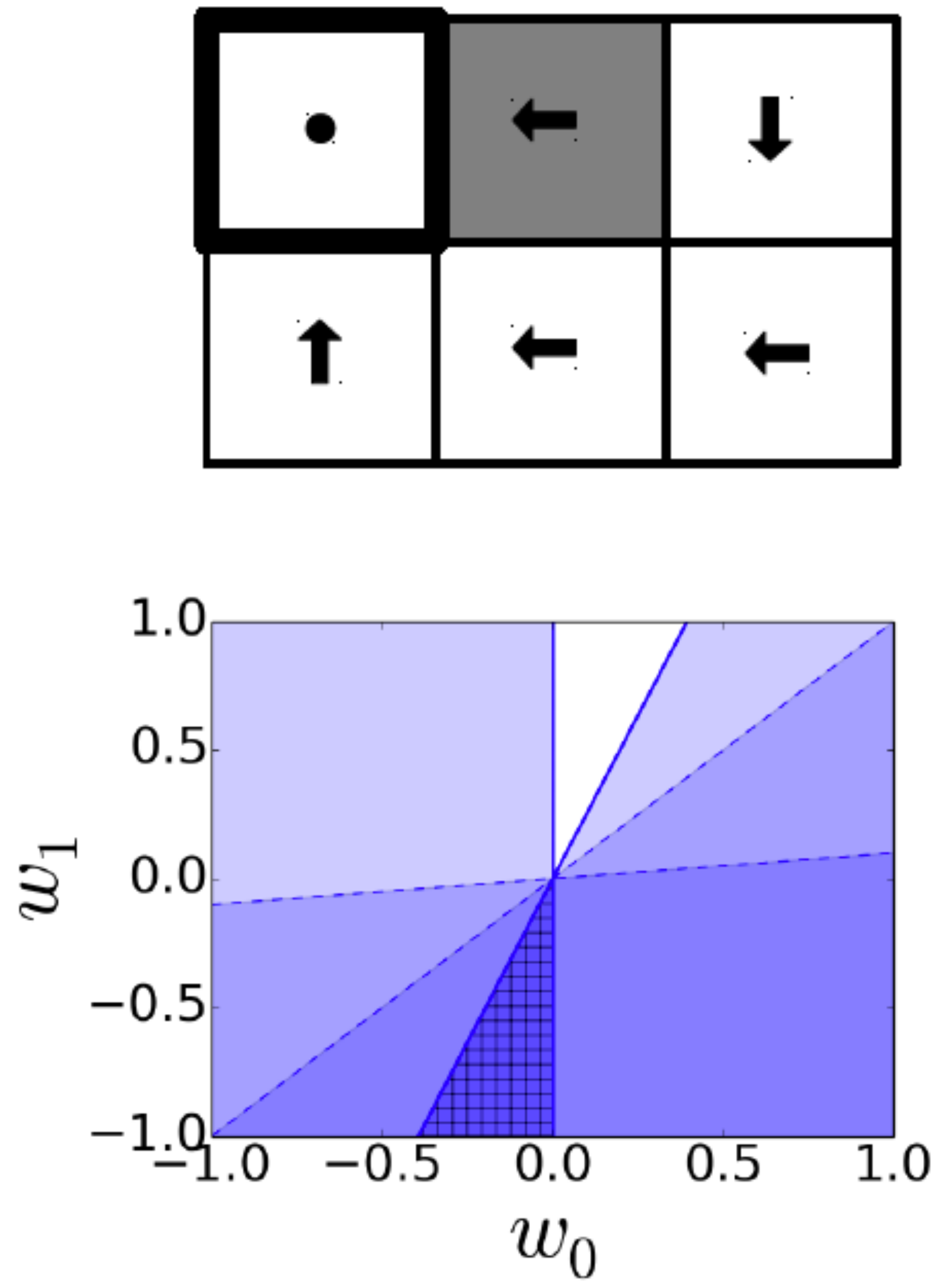


Corollary 1. *BEC($\mathcal{D}|\pi$) is given by the following intersection of half-spaces:*

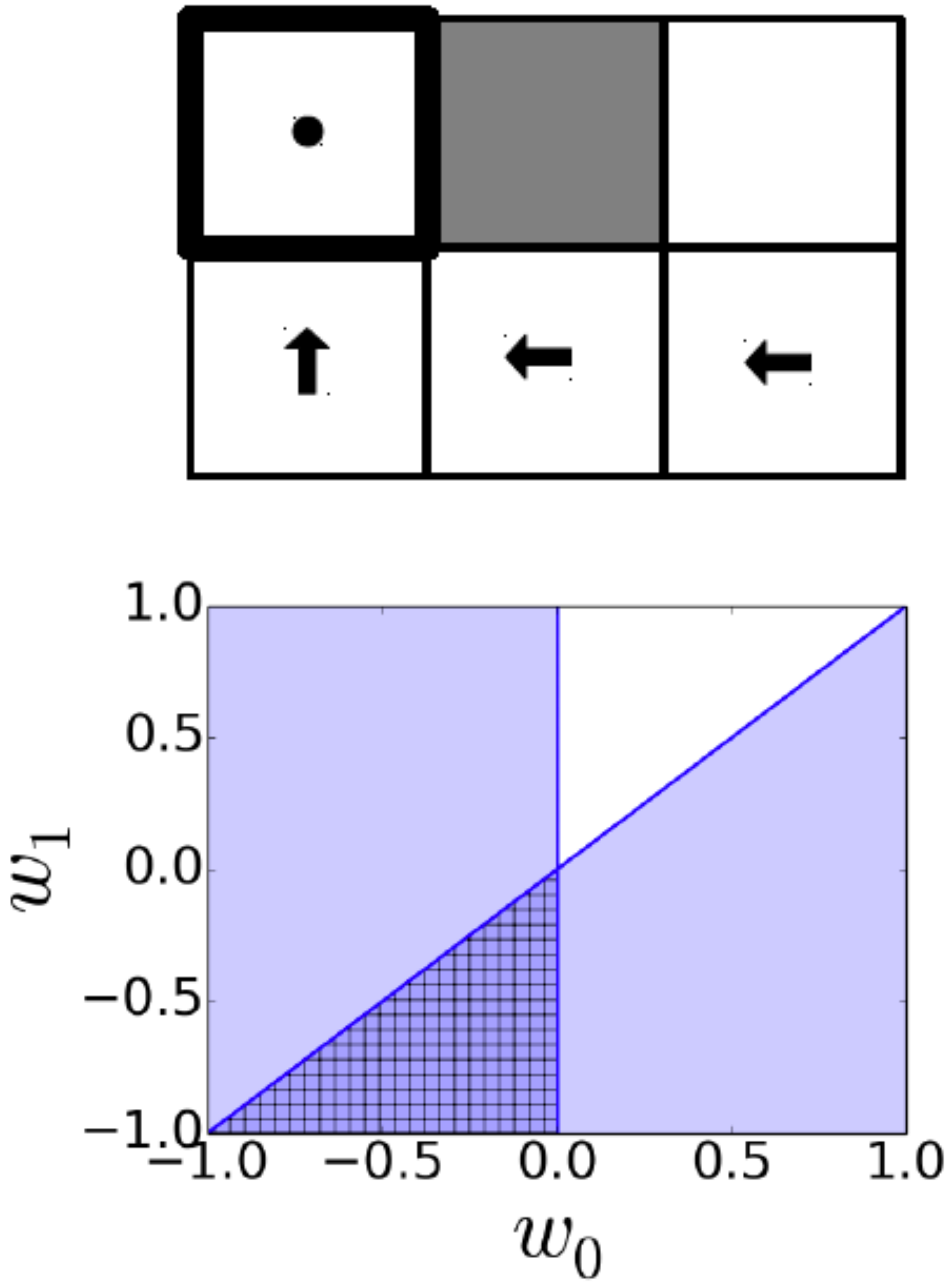
$$\mathbf{w}^T(\mu_\pi^{(s,a)} - \mu_\pi^{(s,b)}) \geq 0, \ \forall (s, a) \in \mathcal{D}, b \in \mathcal{A}.$$

Set Cover Optimal Teaching (SCOT)

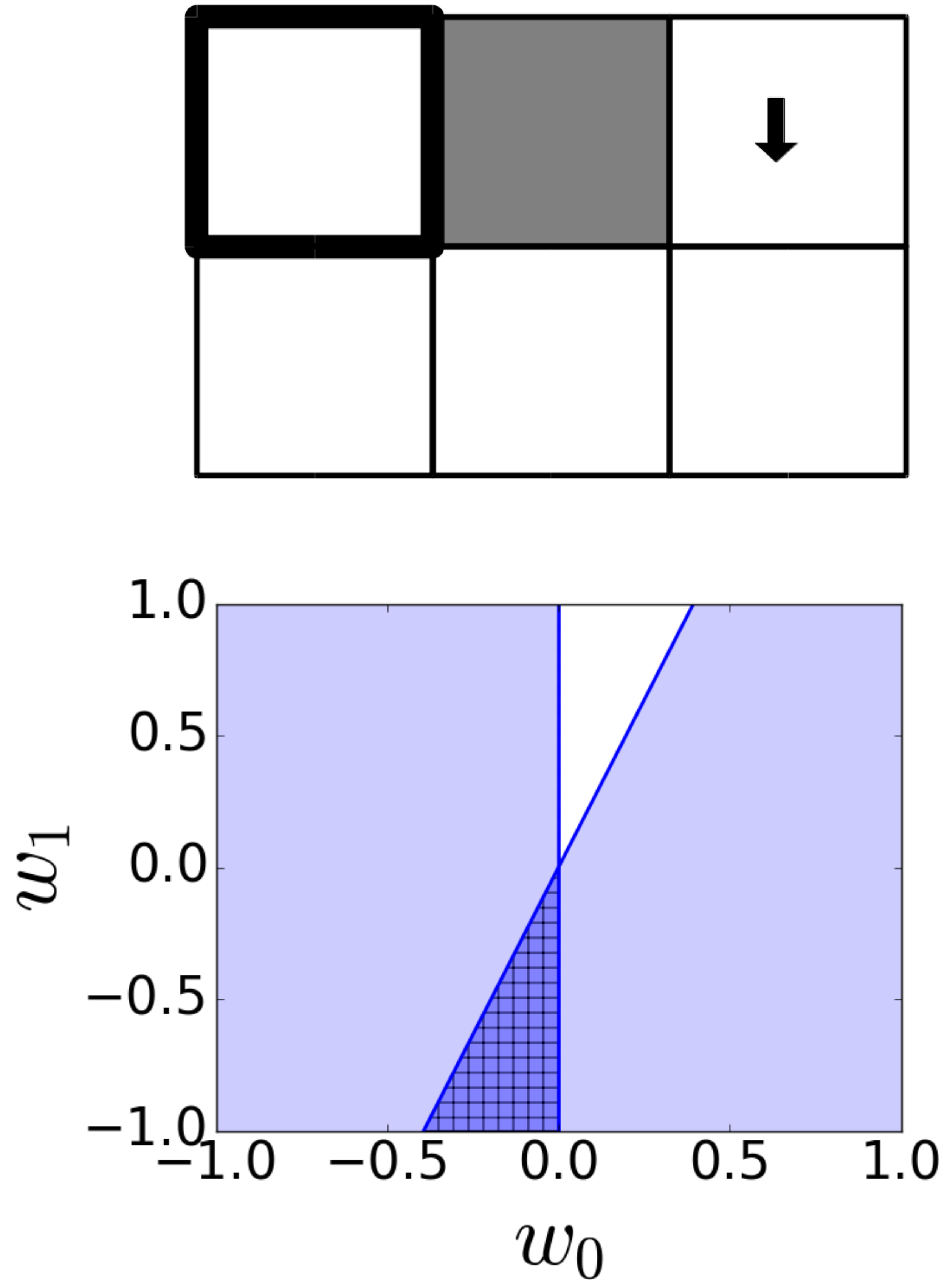
Over-complete



Under-complete



Info-optimal



Submodular = greedy algo approximate optimal!

Information-optimal teaching efficiency

vs. [Cakmak and Lopes 2012]

	Ave. number of (s, a) pairs	Ave. policy loss	Ave. % incorrect actions	Ave. time (s)
UVM (10^5)	5.150	1.539	31.420	567.961
UVM (10^6)	6.650	1.076	19.568	1620.578
UVM (10^7)	8.450	0.555	18.642	10291.365
SCOT	17.160	0.001	0.667	0.965

More accurate AND several orders of magnitude more efficient

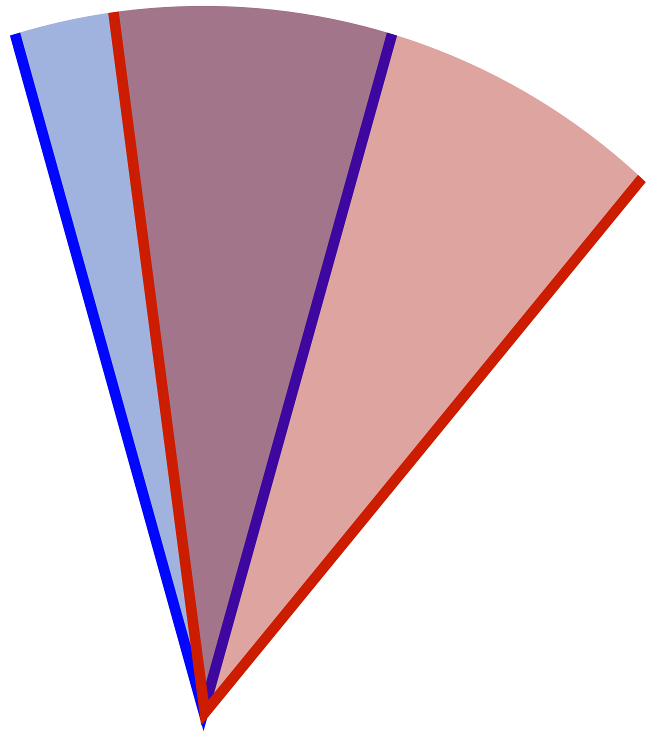
Bayesian Info-Optimal Inverse Reinforcement Learning (BIO-IRL)

$$P(D|R) \propto P_{\text{info}}(\mathcal{D}|R) \cdot \prod_{(s,a)\in\mathcal{D}} P((s,a)|R)$$

$$P_{\text{info}}(\mathcal{D}|R) \propto \exp(-\lambda \cdot \text{infoGap}(\mathcal{D}, R))$$

Prefer rewards that imply expert is both behaviorally optimal and (approximately) information-optimal

Bayesian Info-Optimal Inverse Reinforcement Learning (BIO-IRL)



N-demo remaining volume

N-optimal remaining volume

Intersection of volumes

$$P(D|R) \propto P_{\text{info}}(\mathcal{D}|R) \cdot \prod_{(s,a)\in\mathcal{D}} P((s,a)|R)$$

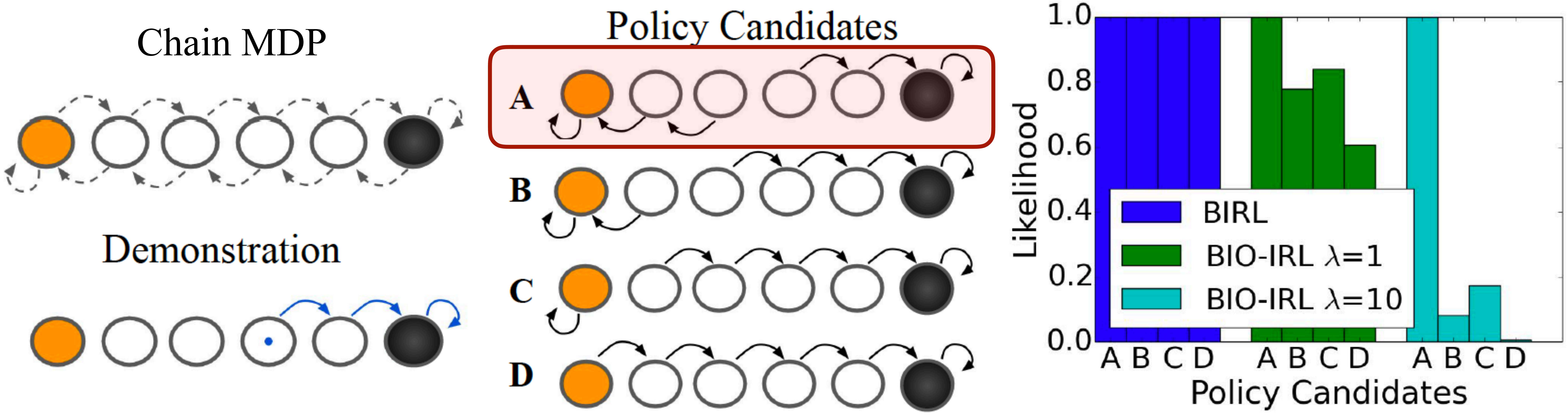
$$P_{\text{info}}(\mathcal{D}|R) \propto \exp(-\lambda \cdot \text{infoGap}(\mathcal{D}, R))$$

Ideally: purple / (red + blue)

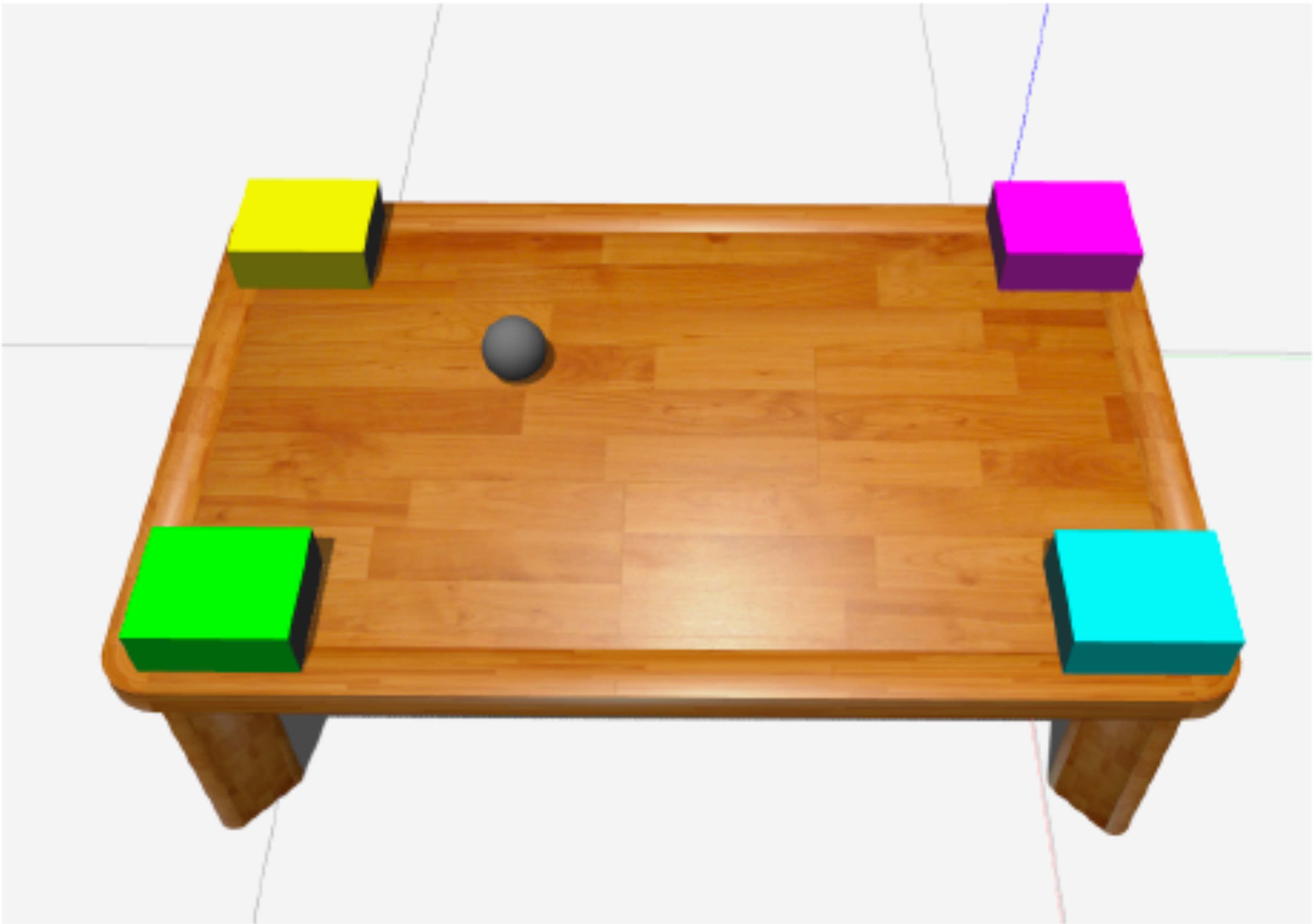
Approx: greedy hyperplane matching + angular distance

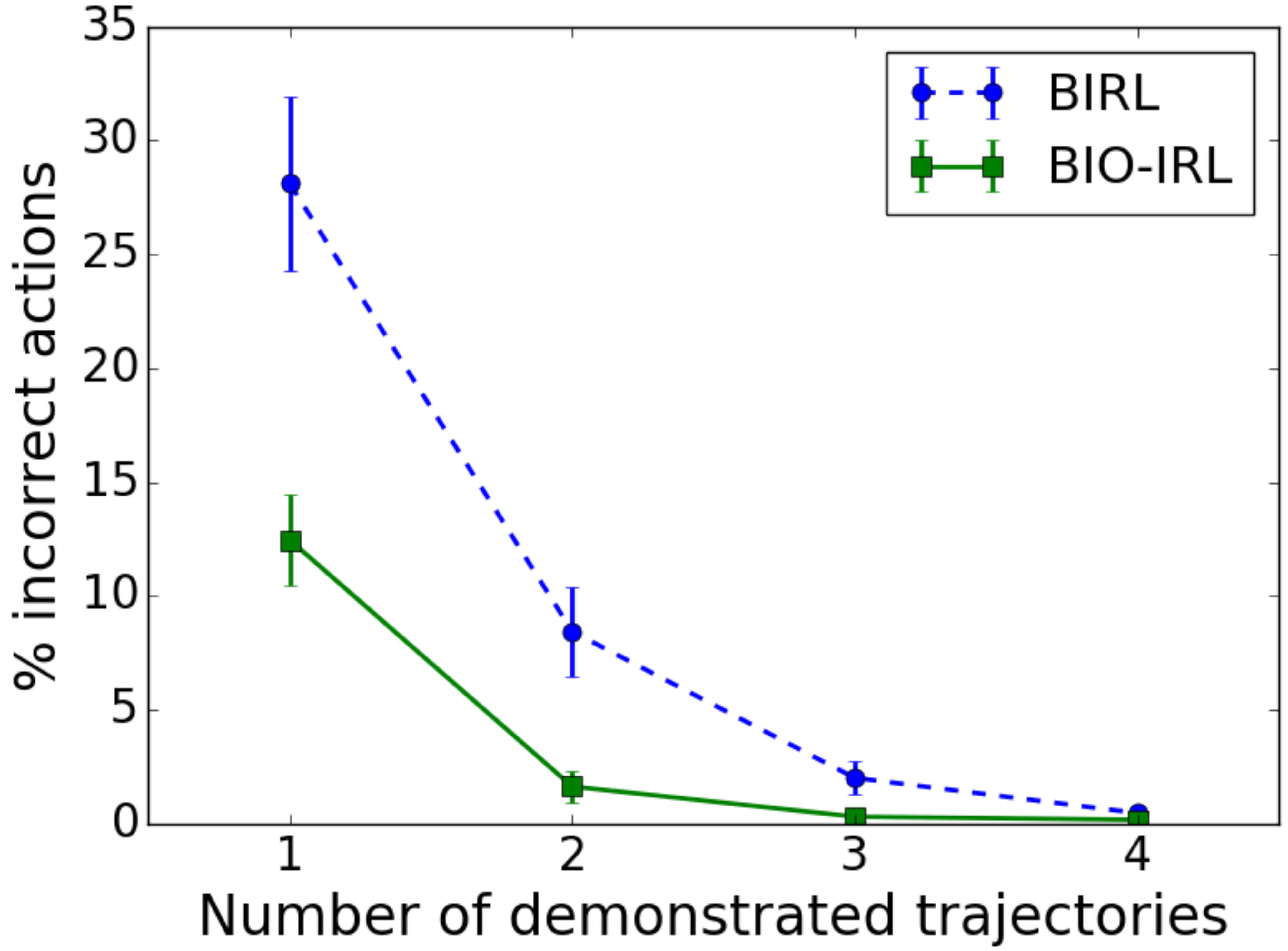
Prefer rewards that imply expert is both behaviorally optimal and (approximately) information-optimal

Example results: I.I.D. vs. information-optimality assumptions



Efficiency gain: I.I.D. vs. information-optimality assumptions





Summary

Re-evaluating bad assumptions for efficient safe RL and imitation learning

- When performing policy evaluation, it is better to collect on-policy data than off-policy data
- Biased models lead to biased estimators
- Worst-case reasoning is the best we can do if we don't know the ground-truth reward function
- Demonstration data should be treated as I.I.D.