CS 690: Human-Centric Machine Learning

Prof. Scott Niekum

Active reward learning and teaching

Difficulties in standard Inverse Reinforcement Learning

- No demonstrations of “bad” actions
- Therefore, difficult to discriminate between actions that are *bad* and actions that were simply *not demonstrated*
- Demonstrations may be *optimal* (from the optimal policy) without being *informative*
- Human may not give informative demonstrations since they don’t know what the robot already knows / doesn’t know or how its learning algorithm works

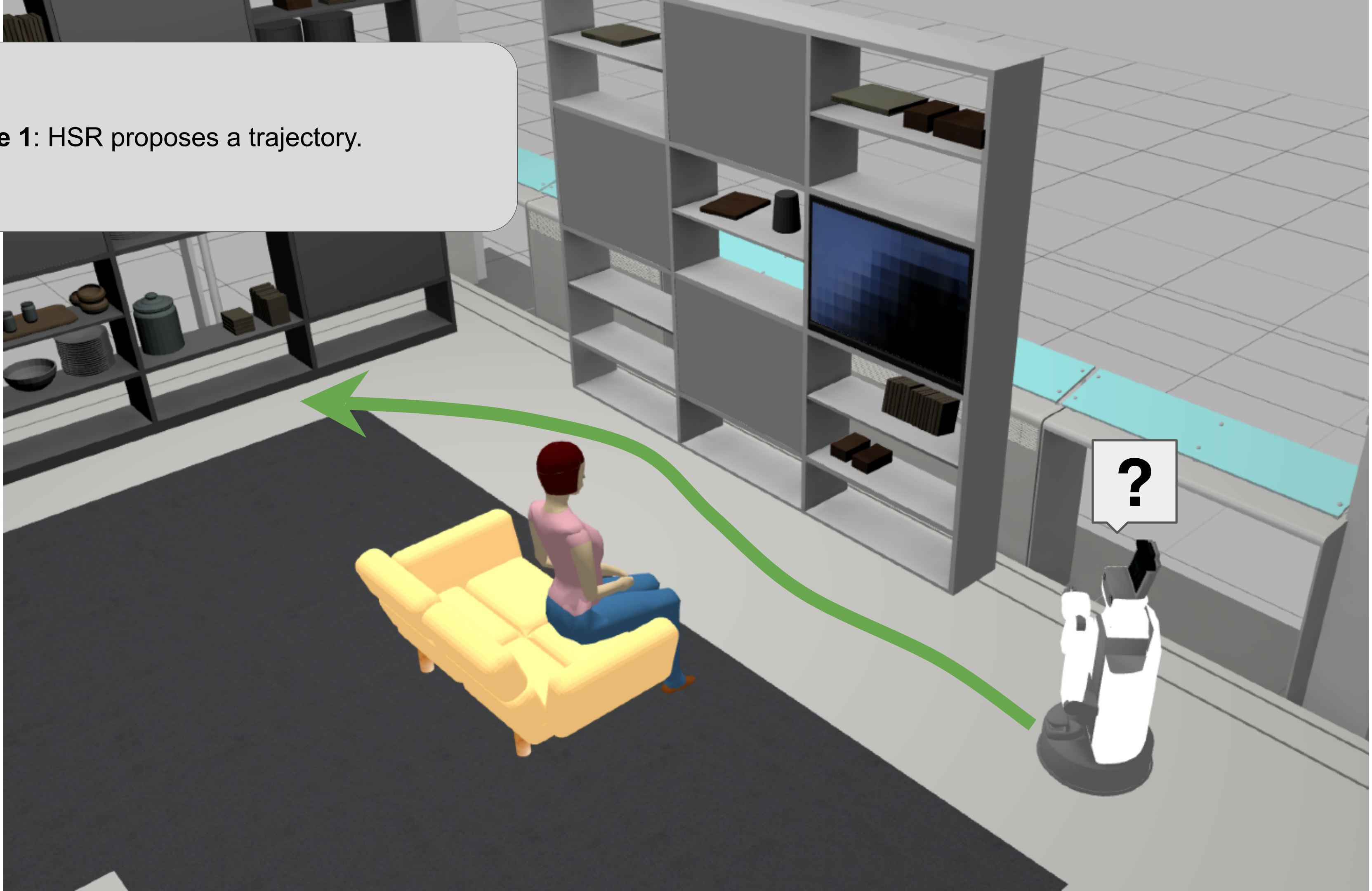
Solution: Active Inverse Reinforcement Learning

- Robot uses knowledge of its current beliefs to generate a query trajectory that will elicit optimally informative feedback from the human (in expectation)
- Human segments the trajectory into *good* and *bad* segments
- Robot updates its beliefs accordingly
- Can be significantly more efficient than requesting a demo from user, can provide direct knowledge about “bad” situations, and requires little effort from human

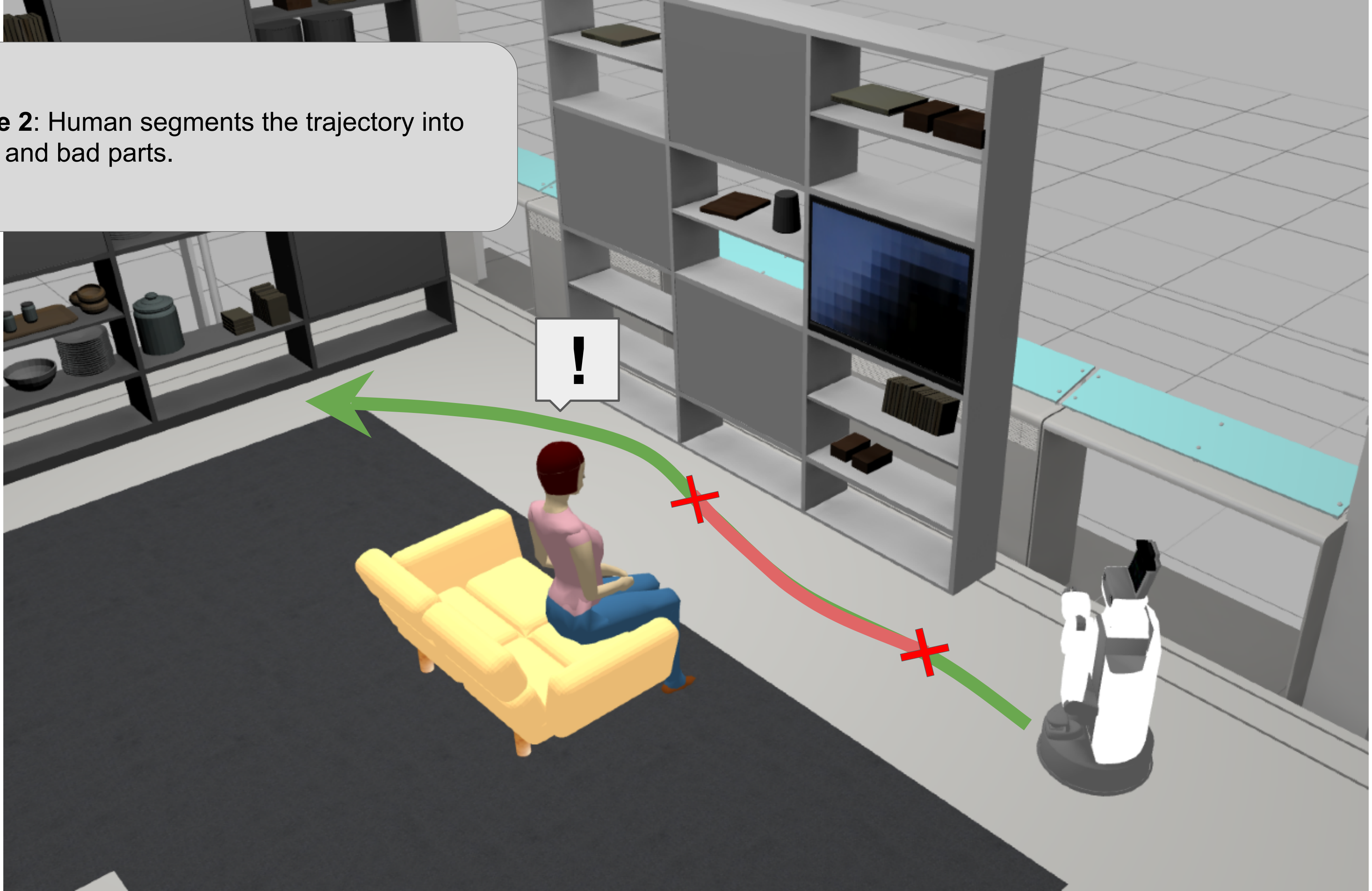
Cui Y, Niekum S. Active reward learning from critiques. IEEE international conference on robotics and automation (ICRA). 2018.

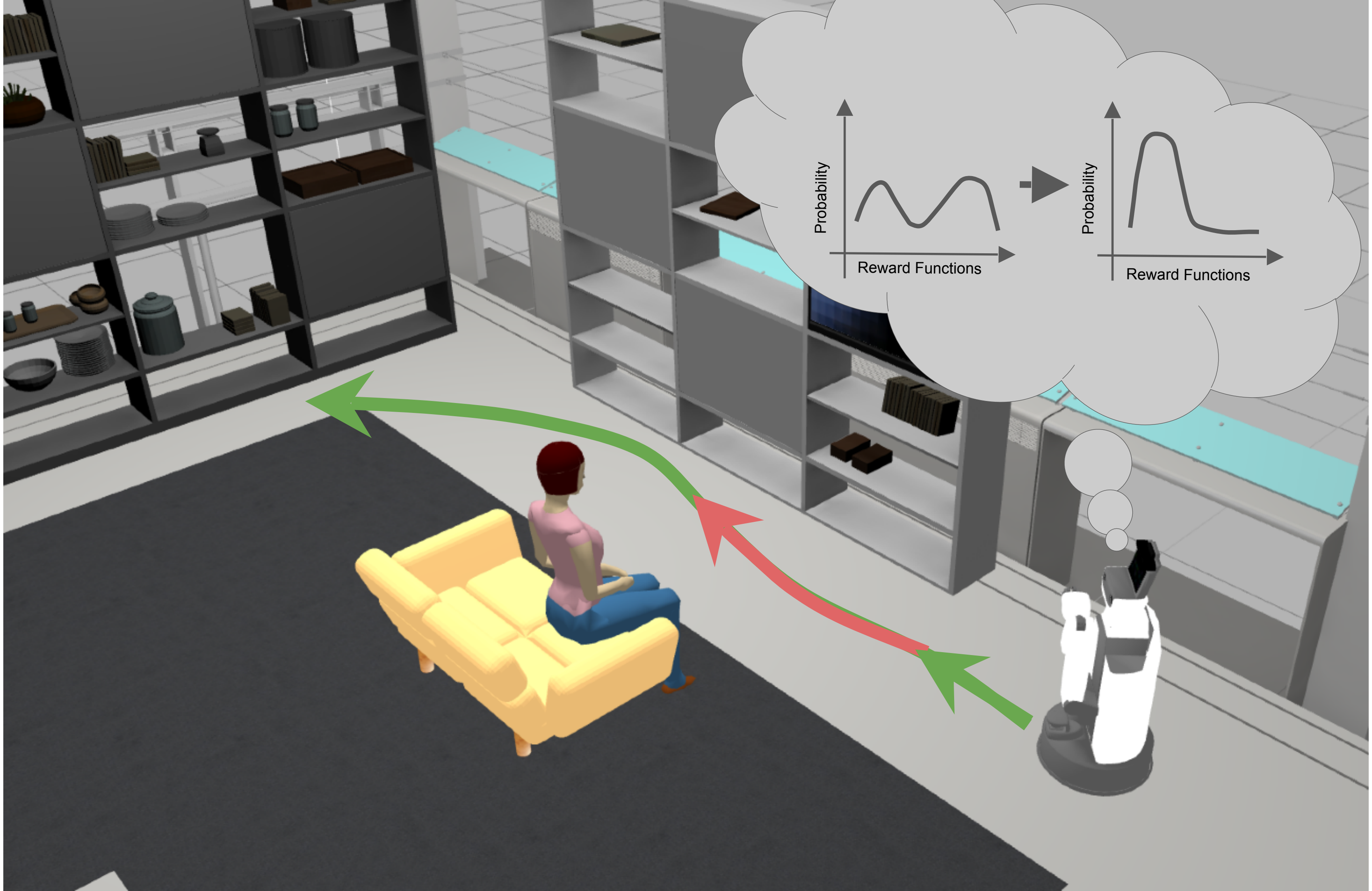
Scenario: HSR is learning how to navigate to the shelf without interrupting the human that is watching TV.

Stage 1: HSR proposes a trajectory.

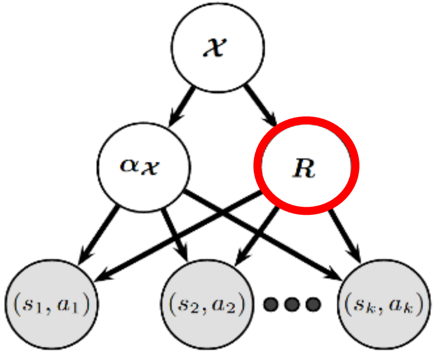


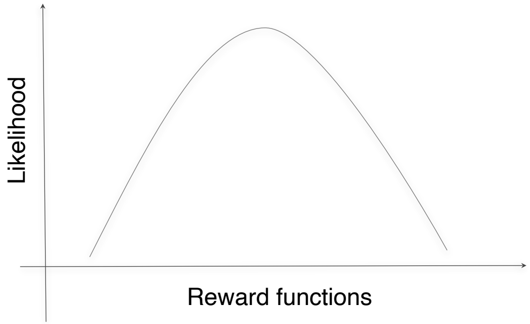
Stage 2: Human segments the trajectory into good and bad parts.




Probability
Reward Functions
Probability
Reward Functions

Bayesian Inverse Reinforcement Learning



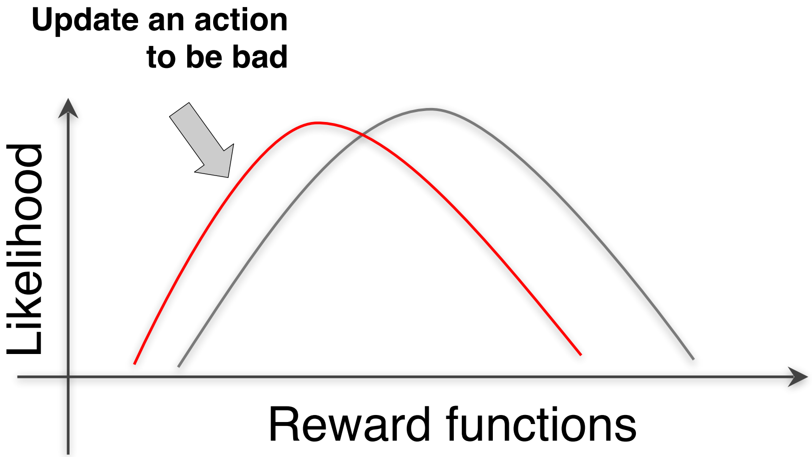


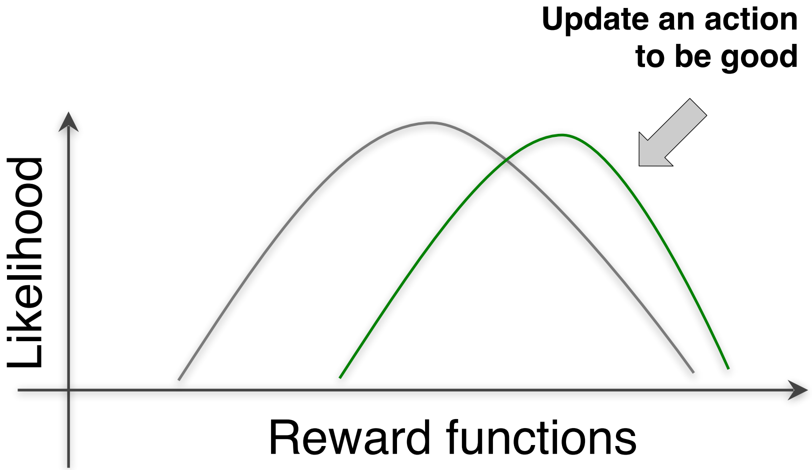
Ramachandran, D., & Amir, E. (2007). Bayesian inverse reinforcement learning. *Urbana*, *51*(61801), 1-4.

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

- Set of optimal actions at a state:

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

- Distance Measure:

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

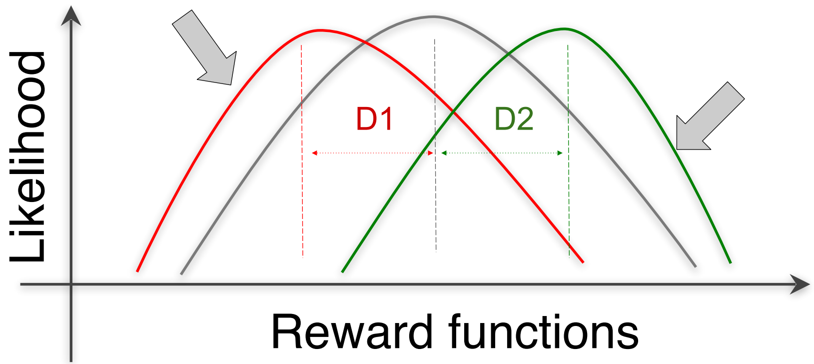
- Expected Information Gain:

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

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

Update an action to be bad

Update an action to be good



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

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

Single (s,a) queries

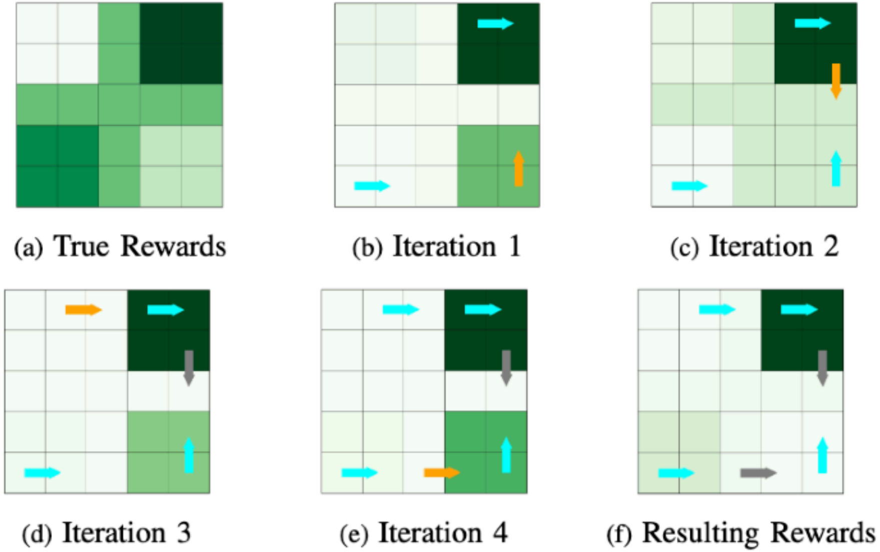
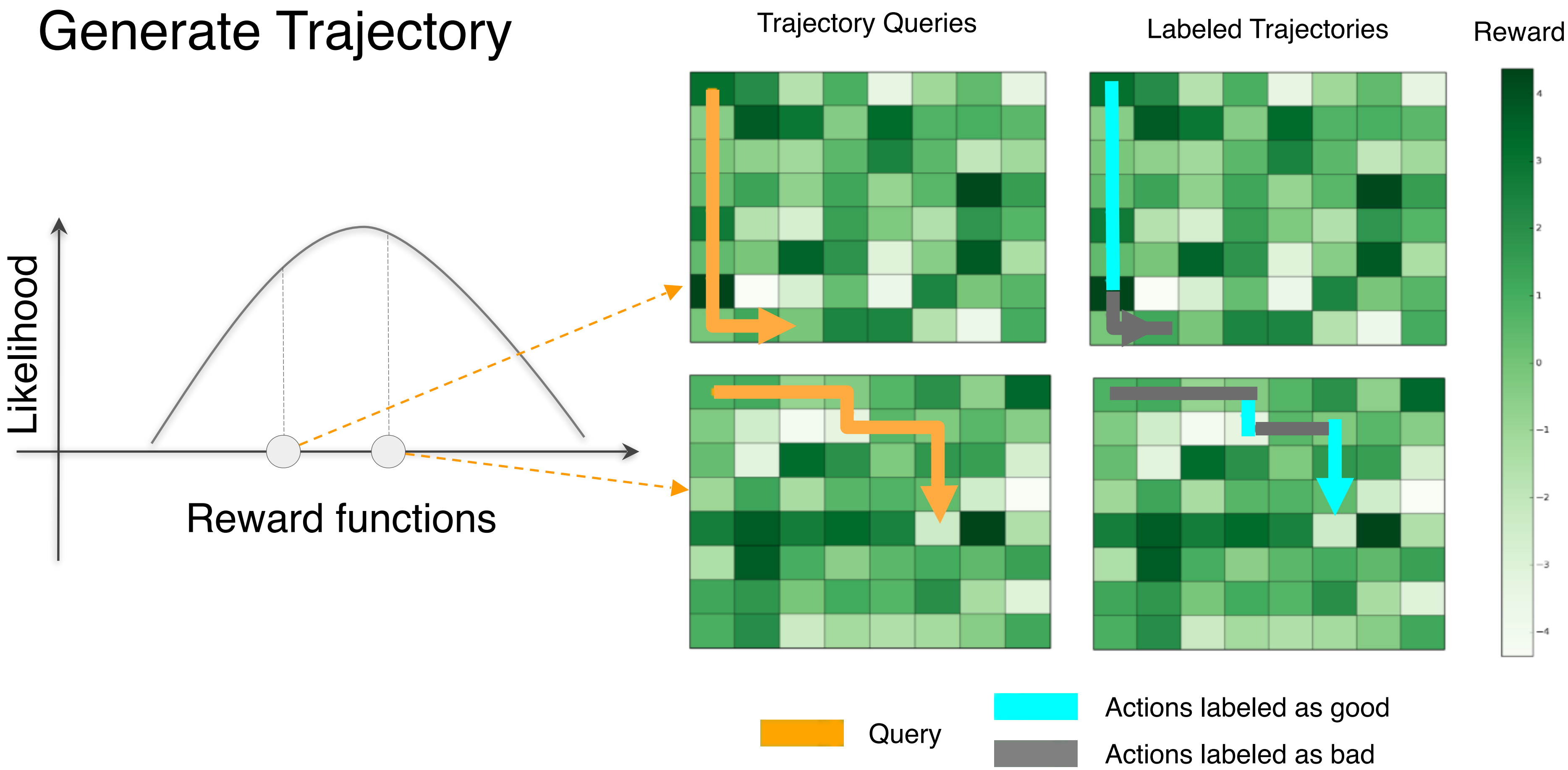
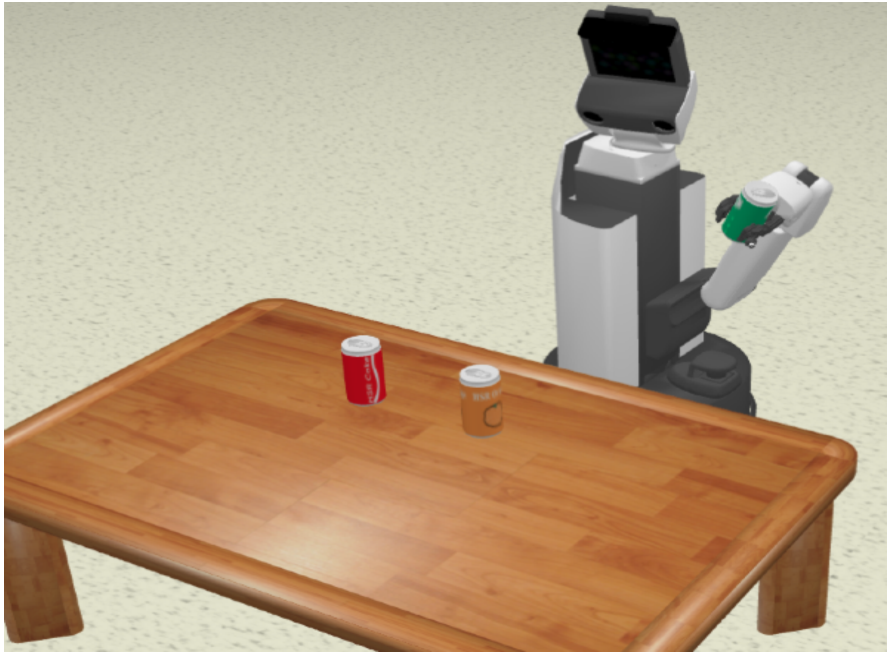


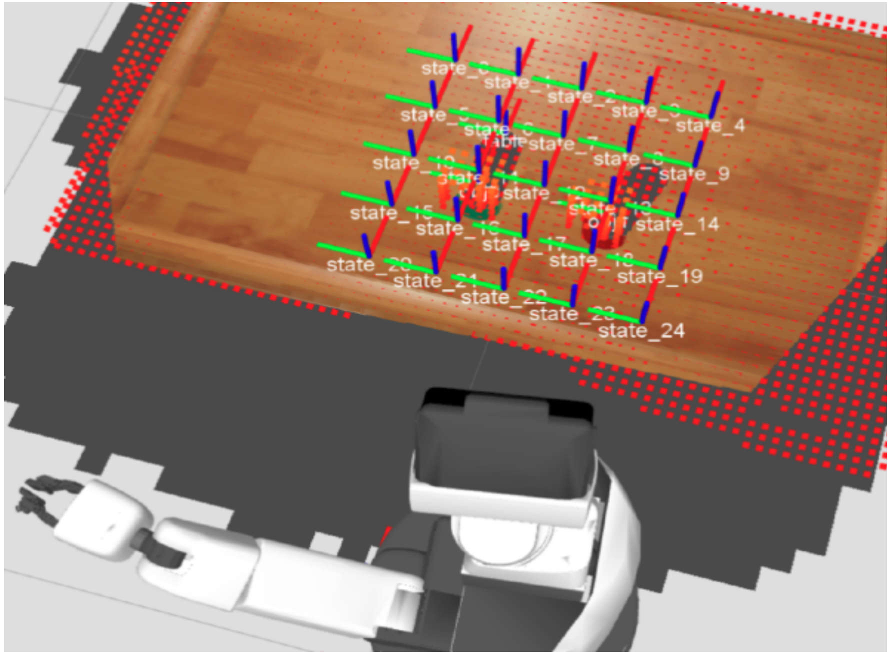
Fig. 2: An illustrative example in a 5×5 gridworld demonstrating actions with maximum expected information gain explore unseen features. Each grid cell has only one of the 5 features. (green: average rewards - darker is larger; cyan: known good actions; gray: known bad actions; orange: actions with max expected info gain)

Iteration	Expected Information Gain	Entropy	Policy Loss
0	-	-	60%
1	4.2753338603	231.58	32%
2	4.2614594772	159.88	28%
3	4.9553412646	151.70	24%
4	5.2887902710	150.42	0%

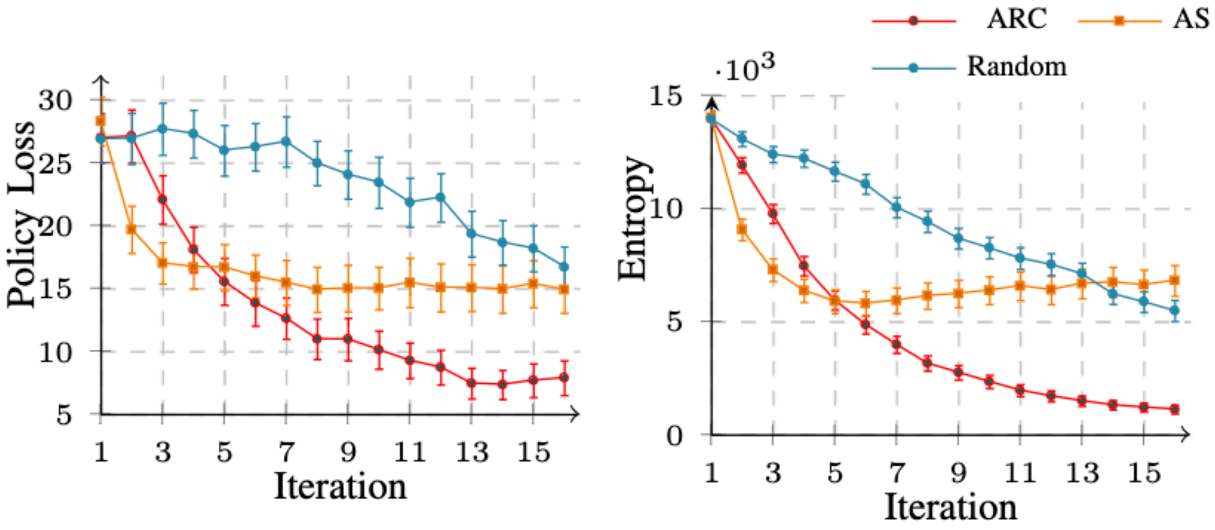

Generate Trajectory
Likelihood
Reward functions
Trajectory Queries
Labeled Trajectories
Reward
4
3
2
1
0
−1
−2
−3
−4
Query
Actions labeled as good
Actions labeled as bad

Task: place an object relative to two objects on a tabletop





Results: Active IRL Policy Loss

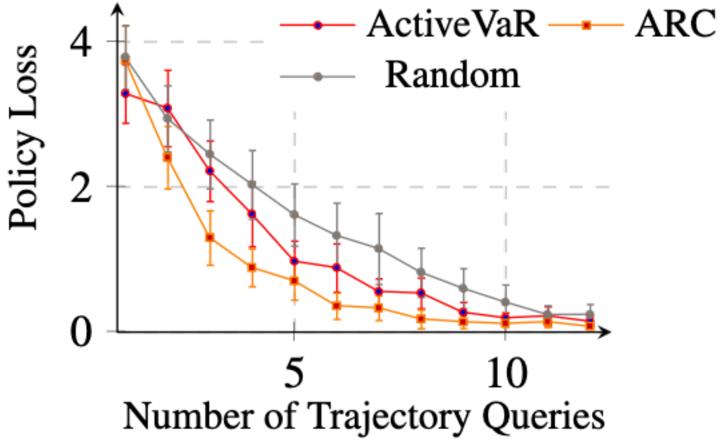


(a) Policy Loss

(b) Entropy over Rewards

Fig. 9: Average Performance on Place-An-Object Task

Alternative: actively improve VaR instead of info gain



(a) Averaged policy losses

Algorithm	Avg. Time (s)
Random	0.0015
ActiveVaR	0.0101
ARC	865.6993

(b) Timing for one iteration of each algorithm

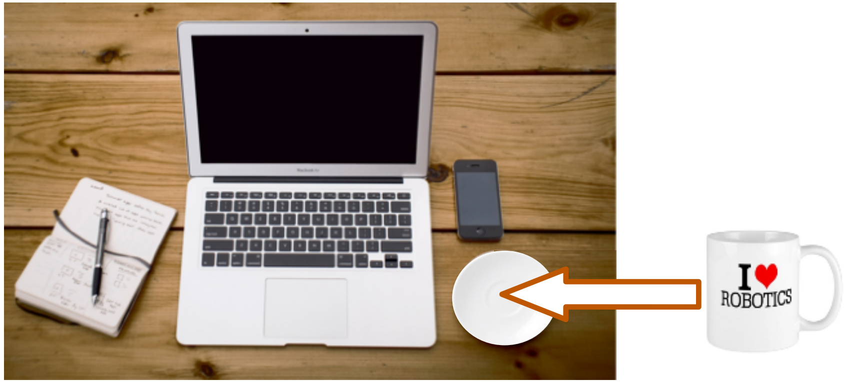
Figure 3: Active critique queries in 8×8 gridworlds with 48 features.

Brown DS, Cui Y, Niekum S. Risk-aware active inverse reinforcement learning. Conference on Robot Learning (CoRL). 2018.

Is this an instantiation of CIRL?

What might this look like for preferences?

Informative demonstrations



Less informative

More informative

Machine teaching

In general:

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

For inverse RL:

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

where:

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

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

Brown DS, Niekum S. Machine teaching for inverse reinforcement learning: Algorithms and applications. Proceedings of the AAAI Conference on Artificial Intelligence. 2019.

Behavioral Equivalence Classes (BEC)

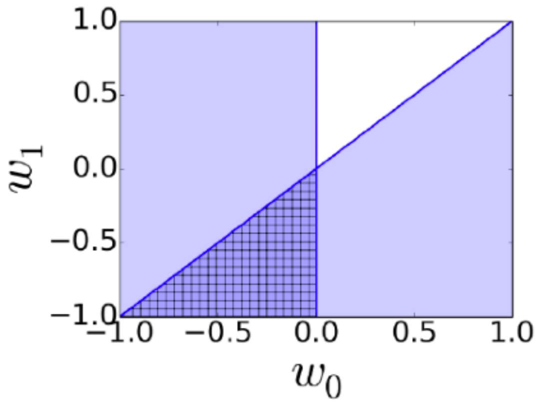
$$\text{BEC}(\pi) =$$
$$\{\mathbf{w} \in \mathbb{R}^k \mid \pi \text{ is optimal under } R(s) = \mathbf{w}^T \phi(s)\}.$$

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

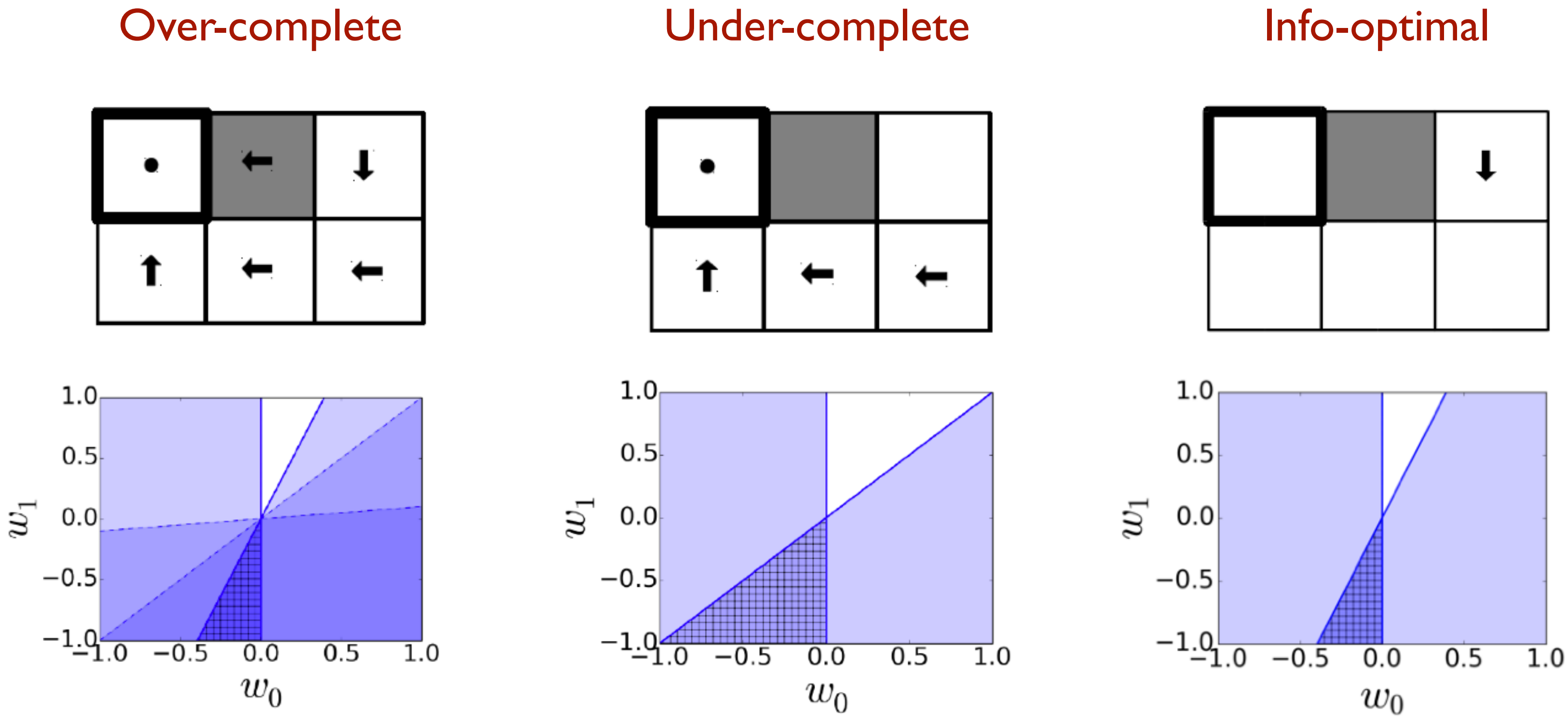
$$\mathbf{w}^T(\mu_\pi^{(s,a)} - \mu_\pi^{(s,b)}) \geq 0,$$
$$\forall a \in \arg\max_{a' \in \mathcal{A}} Q^*(s, a'), b \in \mathcal{A}, s \in \mathcal{S}$$

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)



Submodular = greedy algorithm approximately 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)

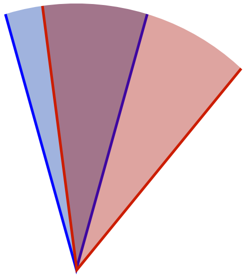
$$P(D|R) \propto P_{\text{info}}(\mathcal{D}|R) \cdot \prod_{(s,a)\in\mathcal{D}} P((s,a)|R)$$

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

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

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

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



N-demo remaining volume

N-optimal remaining volume

Intersection of volumes

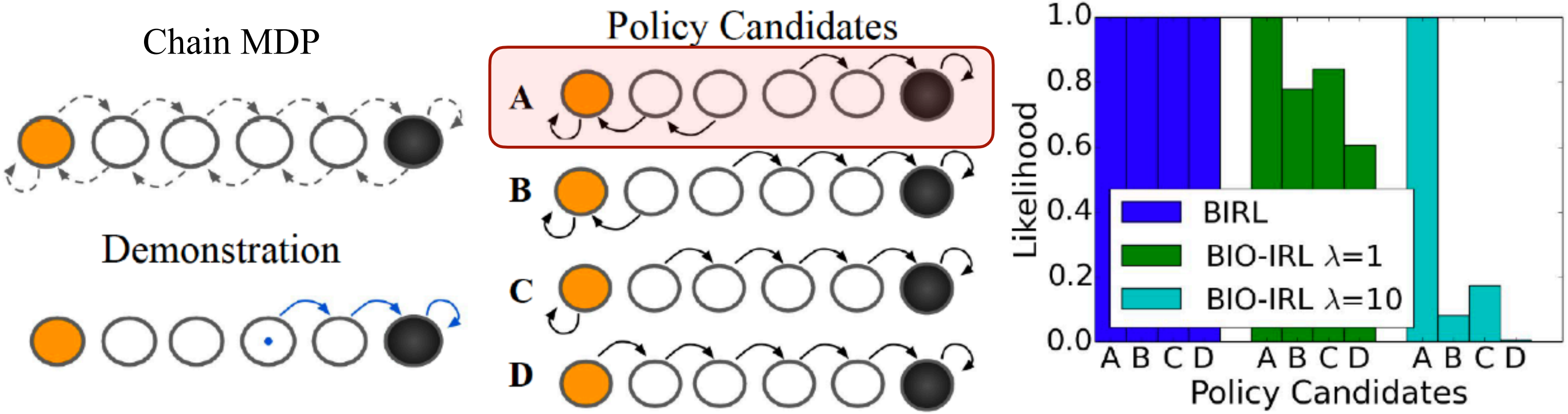
$$P_{\text{info}}(\mathcal{D}|R) \propto \exp(-\lambda \cdot \text{infoGap}(\mathcal{D}, R))$$

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