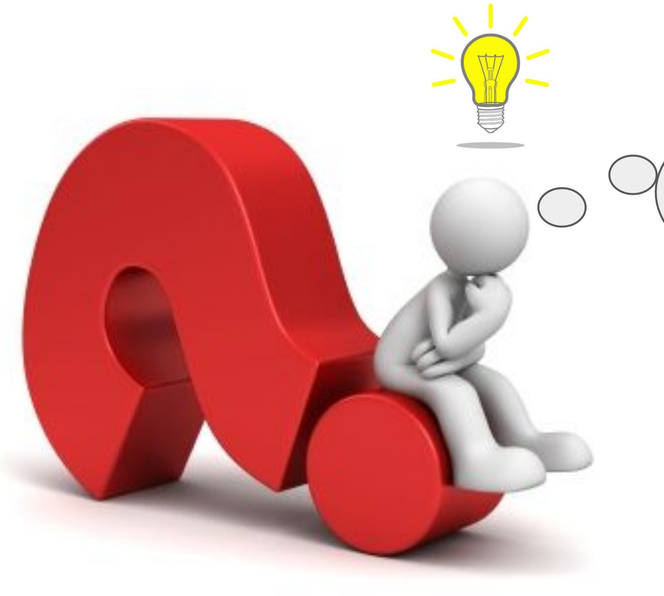
Robust Inverse Reinforcement Learning under Transition Dynamics Mismatch

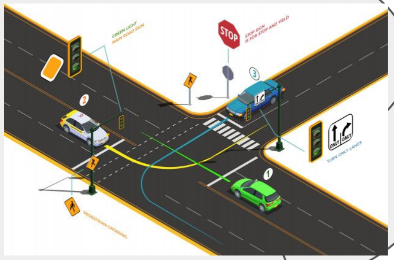
L Viano, YT Huang, P Kamalaruban, A Weller, V Cevher

Why Inverse Reinforcement Learning (IRL)?

Which is the reward function for driving well?



$$R_\theta \, , \, \pi \leftarrow$$



Ziebart et al. 2008, Ziebart 2010

Maximum Causal Entropy IRL (MCE-IRL)

ϕ Are features meant to capture purposeful characteristics of the observed expert behaviour

Linear Reward Function

$$R_{\boldsymbol{\theta}}(s) =< \boldsymbol{\theta}, \phi(s) > \qquad \phi(s) \in \mathbb{R}^d$$

If the transition dynamics are known, than $P[S_t = s \mid \pi, M]$ is known and we can compute:

$$\boldsymbol{\rho}_M^\pi := (1-\gamma) \sum_{s \in \mathcal{S}} \sum_{t=0}^{\infty} \gamma^t P[S_t = s \mid \pi, M] \phi(s)$$

known as occupancy measure

$$V_{M_{\boldsymbol{\theta}}}^\pi = \frac{1}{1-\gamma} < \boldsymbol{\theta}, \boldsymbol{\rho}_M^\pi >$$

$$M_{\boldsymbol{\theta}} = \{\mathcal{S}, \mathcal{A}, T, \gamma, P_0, R_{\boldsymbol{\theta}}\}$$

$$M = M_{\boldsymbol{\theta}} \setminus R_{\boldsymbol{\theta}}$$

Assumptions

- Linear Reward Function : $R_{\boldsymbol{\theta}}(s) =< \boldsymbol{\theta}, \boldsymbol{\phi}(s) >$ $\boldsymbol{\phi}(s) \in \mathbb{R}^d$
- We introduce the occupancy measure as: $\boldsymbol{\rho}_M^\pi := (1-\gamma) \sum_{s \in \mathcal{S}} \sum_{t=0}^{\infty} \gamma^t P[S_t = s \mid \pi, M] \boldsymbol{\phi}(s)$
- The Value function is also linear: $V_{M_{\boldsymbol{\theta}}}^\pi = \frac{1}{1-\gamma} < \boldsymbol{\theta}, \boldsymbol{\rho}_M^\pi >$

For any θ , policies with same occupancy measure achieves the same value function.

Maximum Causal Entropy IRL (MCE-IRL)

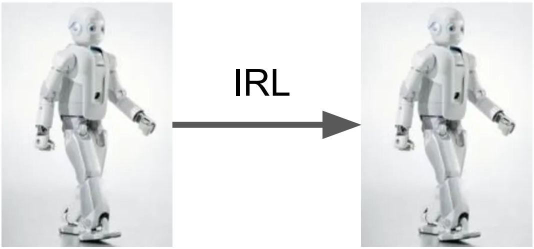
$$V^{\pi}_{M_{\boldsymbol{\theta}}} = \frac{1}{1-\gamma} < \boldsymbol{\theta}, \boldsymbol{\rho}^{\pi}_{M} >$$

$$\text{argmax}_{\pi\in\Pi} \quad \mathbb{E}[\sum_{t=0}^{\infty} -\gamma^{t} \log \pi(a_t|s_t) \Big| \pi, M^{L}]$$

$$\text{subject to} \quad \boldsymbol{\rho}^{\pi}_{M^{L}} = \boldsymbol{\rho}$$

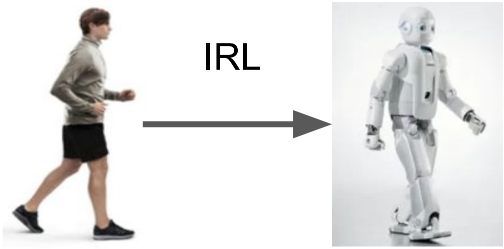
Teacher's Occupancy Measure

Learner's Occupancy Measure

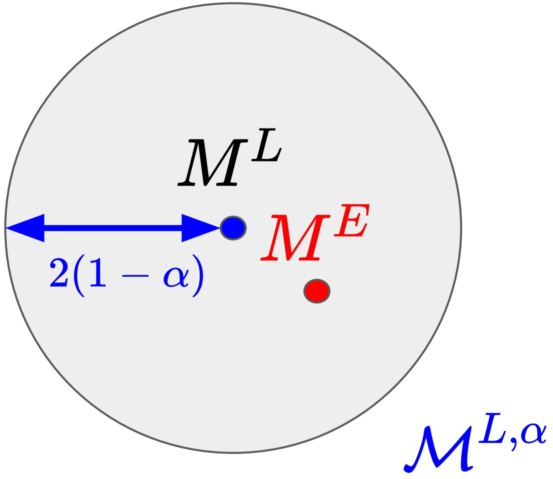


Absence of mismatch

Robust MCE-IRL under transition dynamics mismatch



Mismatch

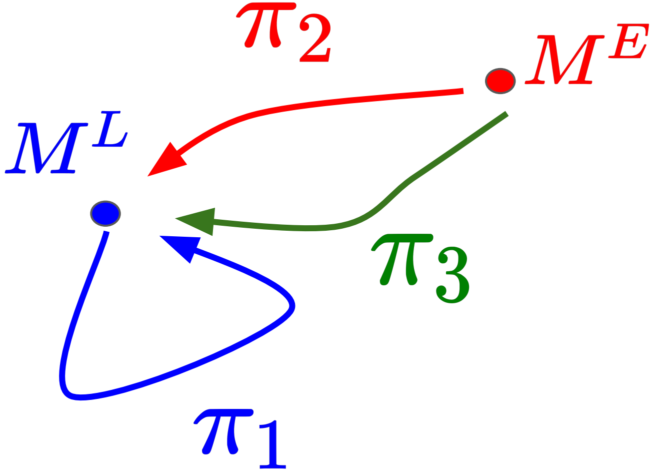


Solve the dual program under the worst case environment **under a fixed** α :

$$\operatorname{argmax}_{\pi \in \Pi} \min_{M \in \mathcal{M}^{L,\alpha}} \mathbb{E}\Big[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t) \,\Big|\, \pi, M\Big]$$

$$\text{subject to} \quad \boldsymbol{\rho} = \boldsymbol{\rho}_M^{\pi}$$

Worst case upper bound for MCE-IRL



$$d(T^E, T^L) := \max_{s,a} ||T^E(\cdot|s,a) - T^L(\cdot|s,a)||_1$$

π_2 **MCE IRL**

π_3 **ROBUST MCE IRL**

$$|V_{T^L}^{\pi_1} - V_{T^L}^{\pi_2}| \leq \mathcal{O}\left(d(T^E, T^L)\right)$$

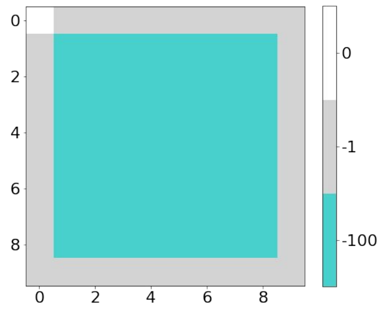
$$|V_{T^L}^{\pi_1} - V_{T^L}^{\pi_3}| \leq \mathcal{O}\left((1+\alpha)d(T^E, T^L) + (1-\alpha)d(T^E, T^*) + 2(1-\alpha)^2\right)$$

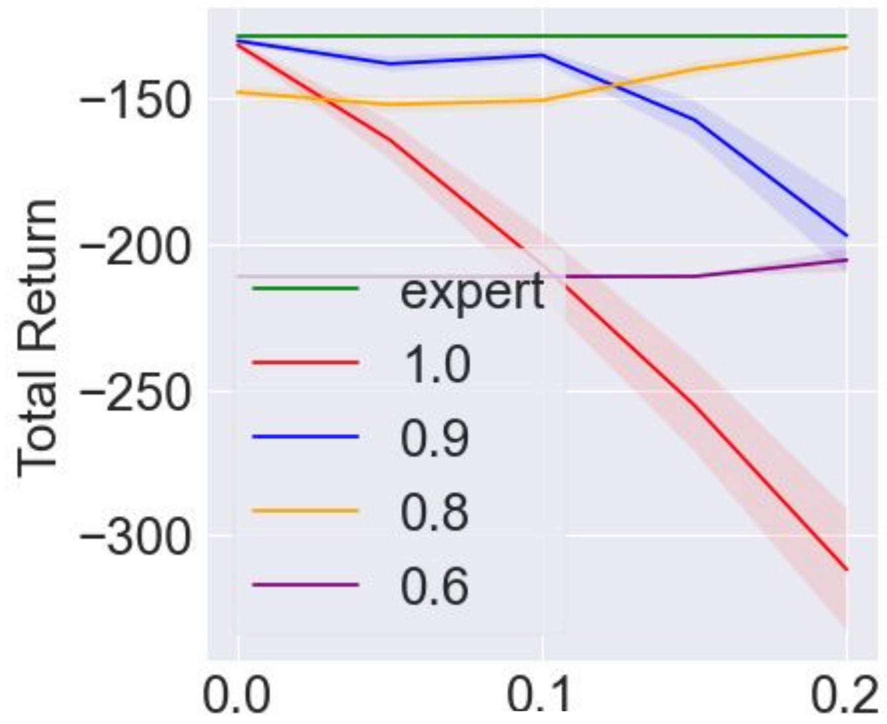
Results with Linear Reward Function

Recall: Standard MCE-IRL (Ziebart, 2010) is recovered with

$$\alpha = 1$$

The legend reports the value for α

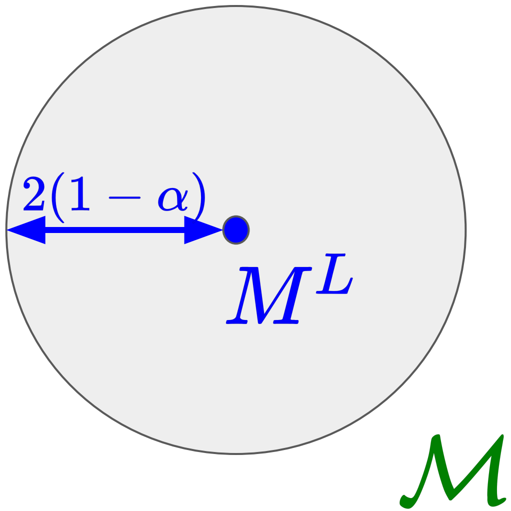




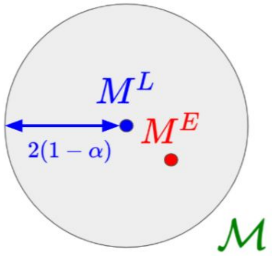
Small Mismatch

Large Mismatch

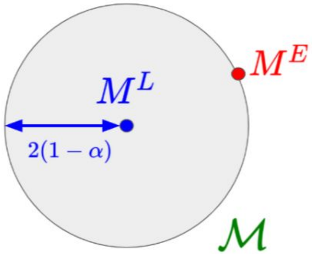
MDP uncertainty region



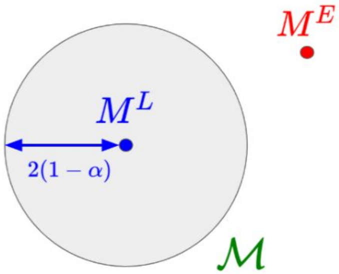
Choosing the right uncertainty set



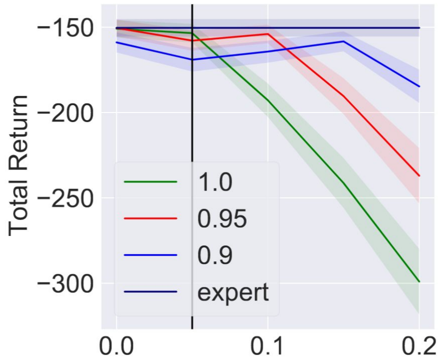
(a) Overestimating $1-\alpha$



(b) Perfect estimation of $1-\alpha$



(c) Underestimating $1-\alpha$

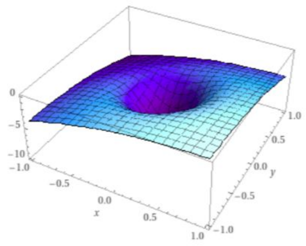


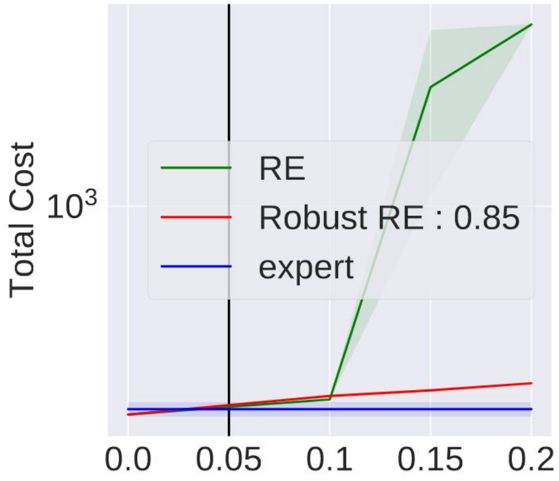
Continuous States and Actions

We propose an extension to continuous states and actions based on Relative Entropy IRL.

Relative Entropy-IRL (Boularias et al., 2011) is recovered with

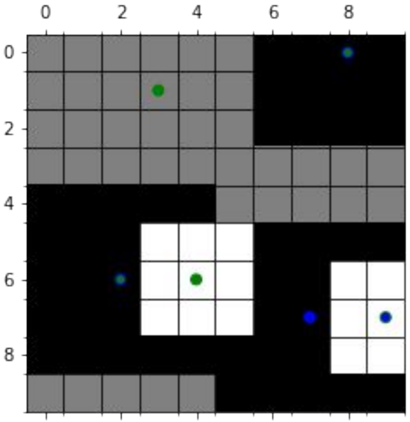
$$\alpha = 1$$

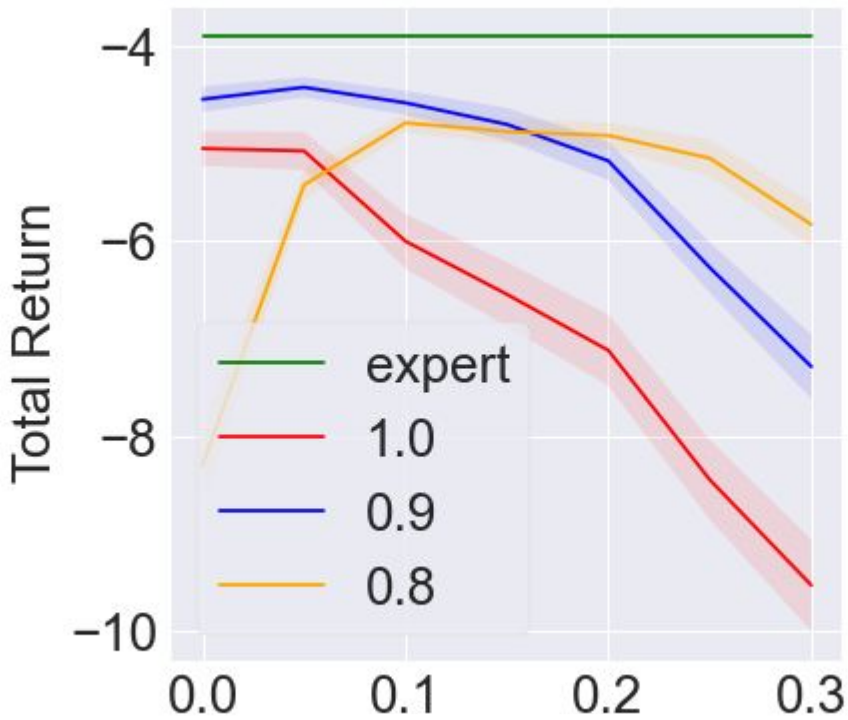




Non Linear Reward Function

Notice: Standard Deep MCE-IRL (Wulfmeier et al. 2015) is recovered with $\alpha = 1$





Small Mismatch

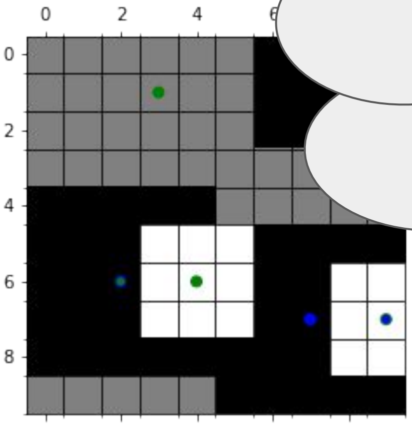
Large Mismatch

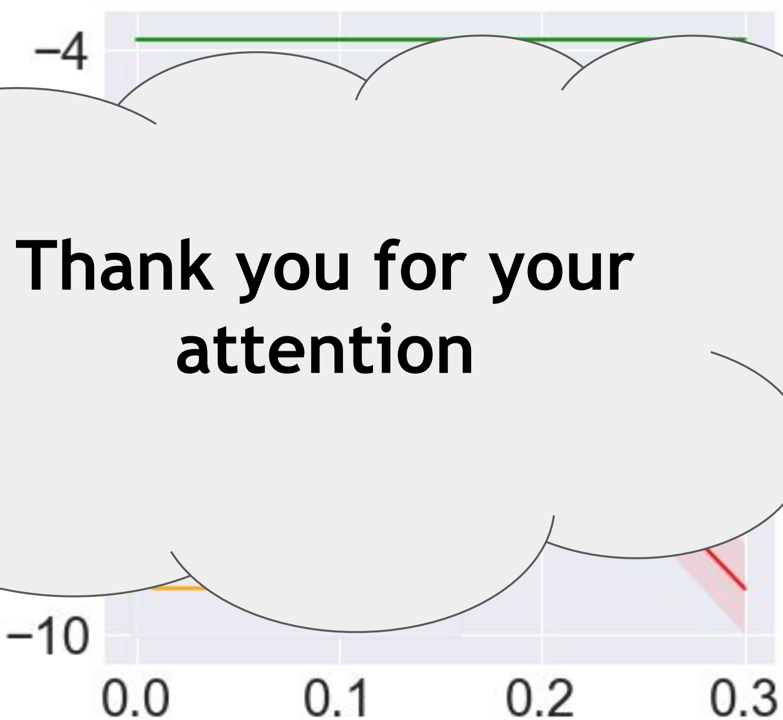
The legend reports the value for α

Non Linear Reward Function

Notice: Standard Deep MCE-IRL (Wulfmeier et al. 2015) is recovered with $\alpha = 1$

The legend reports value for α





Small Mismatch

Large Mismatch

Thank you for your attention