NEURAL INFORMATION PROCESSING SYSTEMS

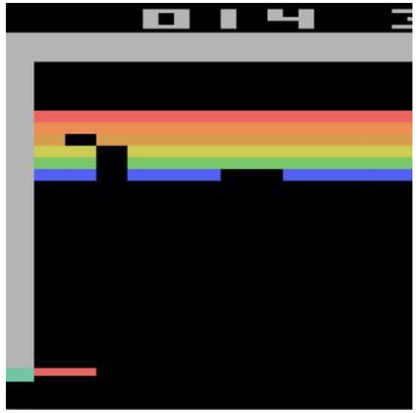
IQ-Learn: Inverse soft-Q learning for Imitation

Divyansh Garg, Shuvam Chakraborty
Chris Cundy, Jiaming Song, Stefano Ermon

Reinforcement Learning

- Goal: Learn policies
- Input: High-dimensional observations

RL needs reward signal



Soft Q-Learning

- Goal: Learn an energy function **Q** to represent the policy

$$\pi(a|s) = \frac{1}{Z_s}\exp(Q(s,a))$$

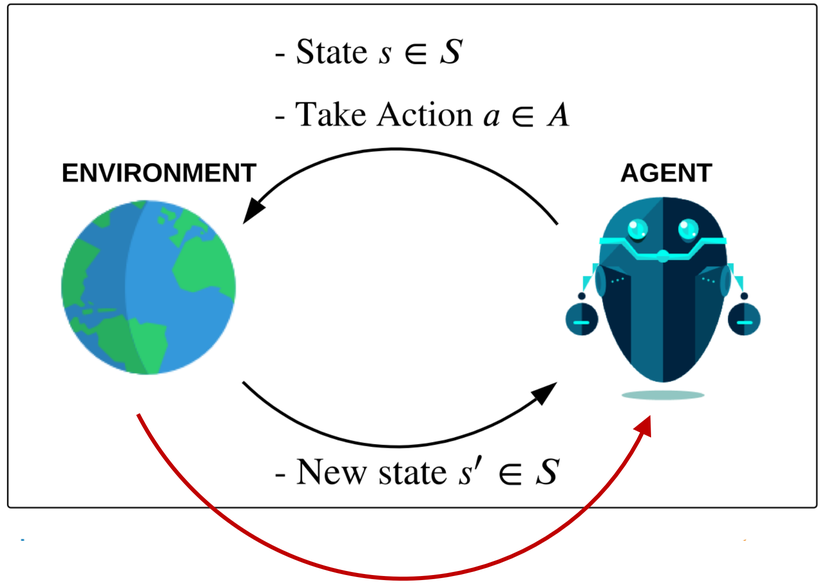
Reward Function (R) ⇨ **Reinforcement Learning (RL)** ⇨ **Q-function (Q)** → **Policy**

Imitation Learning

- Goal: Learn reward function (R) or policy
- Input: Expert demonstrations

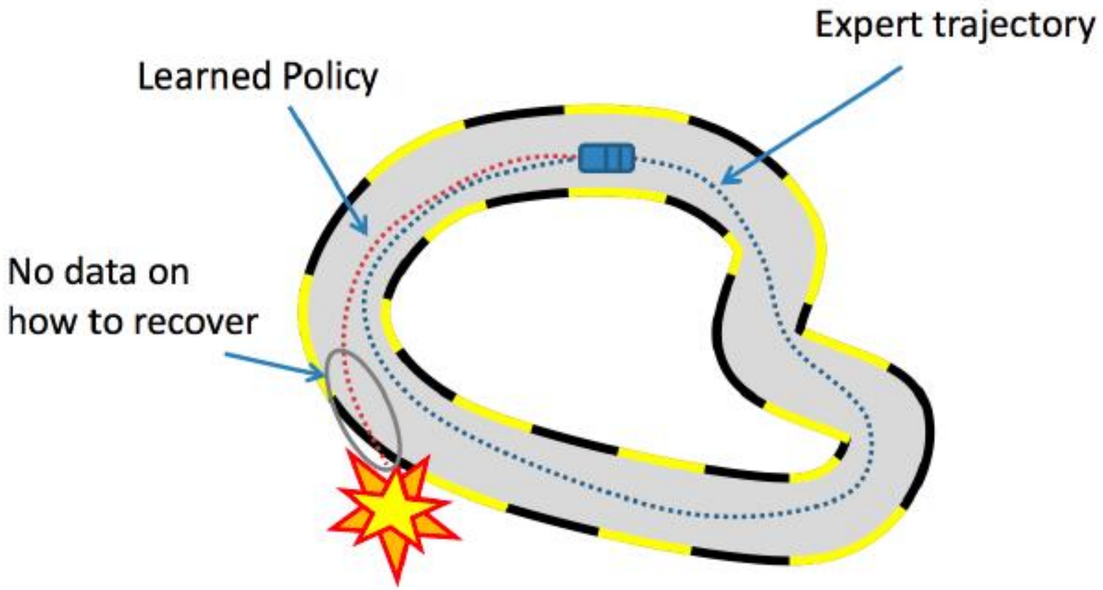
$$\{(s_0^i, a_0^i, s_1^i, a_1^i, \dots)\}_{i=1}^n \sim \pi_E$$

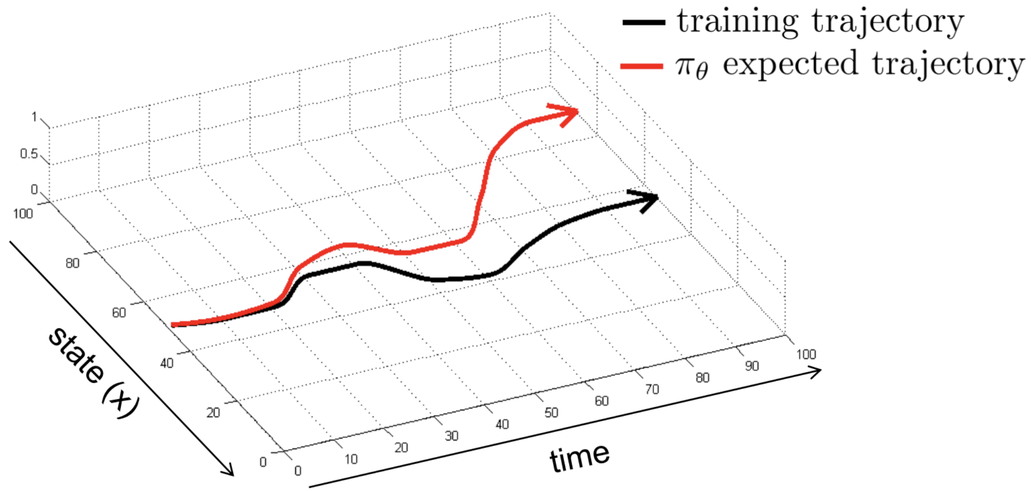
- Input: (Optional) Environment samples



Behavior Cloning

$$\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$$





Inverse Reinforcement Learning

- Goal: Learn reward function (R) or policy
- Input: Expert demonstrations

$$\{(s_0^i, a_0^i, s_1^i, a_1^i, \dots)\}_{i=1}^n \sim \pi_E$$

- Input: (Optional) Environment samples

Learn a reward R such that

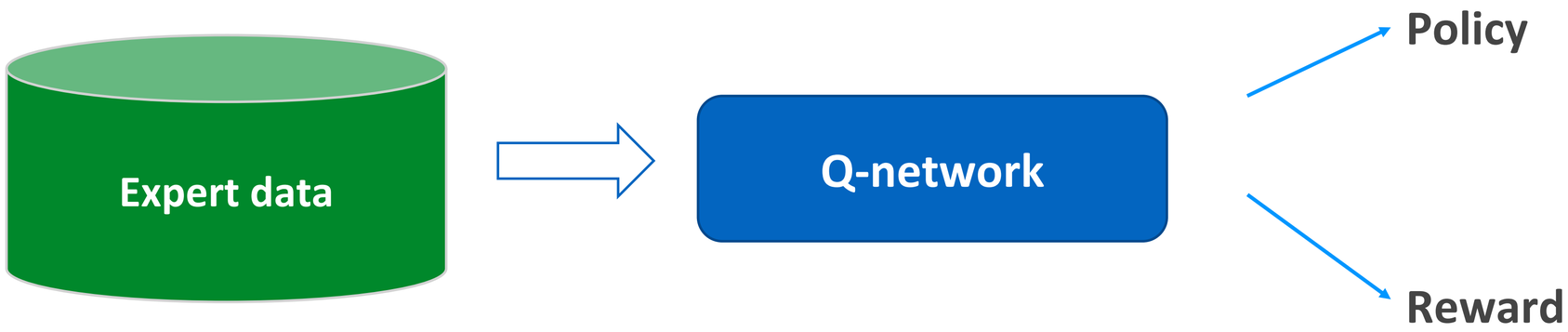
$$\pi_E = \arg\max_\pi \mathbb{E}_\pi[r(s,a)]$$

Inverse Q Learning

- Goal: ~~Learn reward function (R) or policy~~ Learn a Q-function
- Input: Expert demonstrations

$$\{(s_0^i, a_0^i, s_1^i, a_1^i, \dots)\}_{i=1}^n \sim \pi_E$$

- Input: (Optional) Environment samples

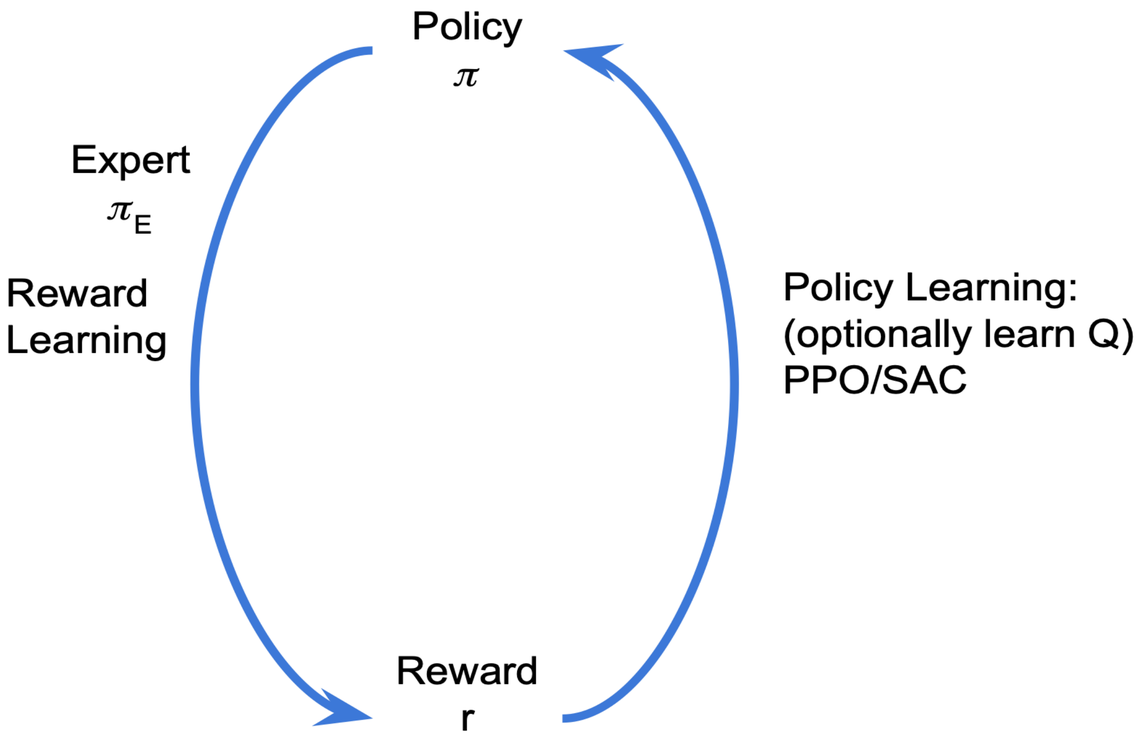


Inverse Q Learning

- Equivalent to Inverse Reinforcement Learning

Much simpler to solve!!!

Inverse Q Learning



Adversarial Learning:
GAIL/AIRL

Results: Offline IL

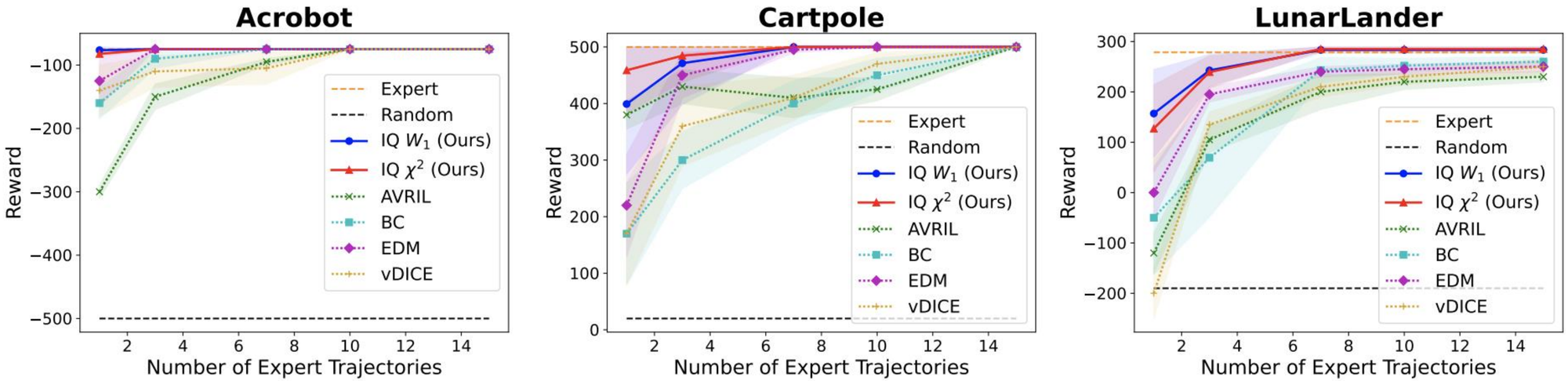
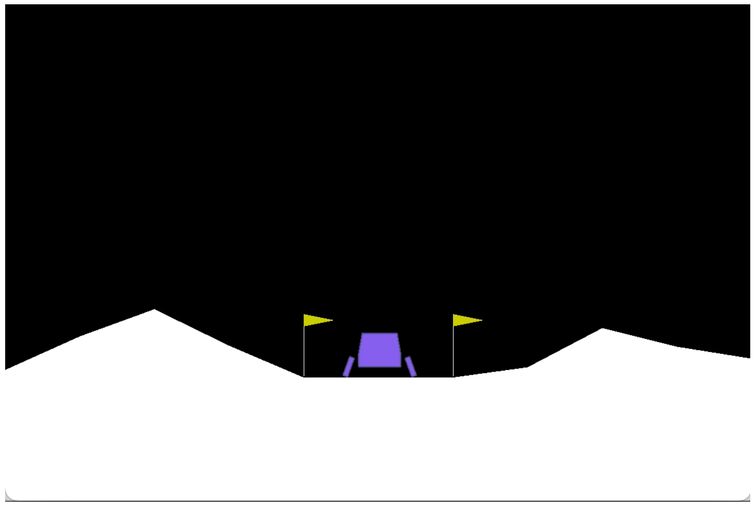


Figure 2: **Offline IL results.** We plot the average environment returns vs the number of expert trajectories.



Results: Atari

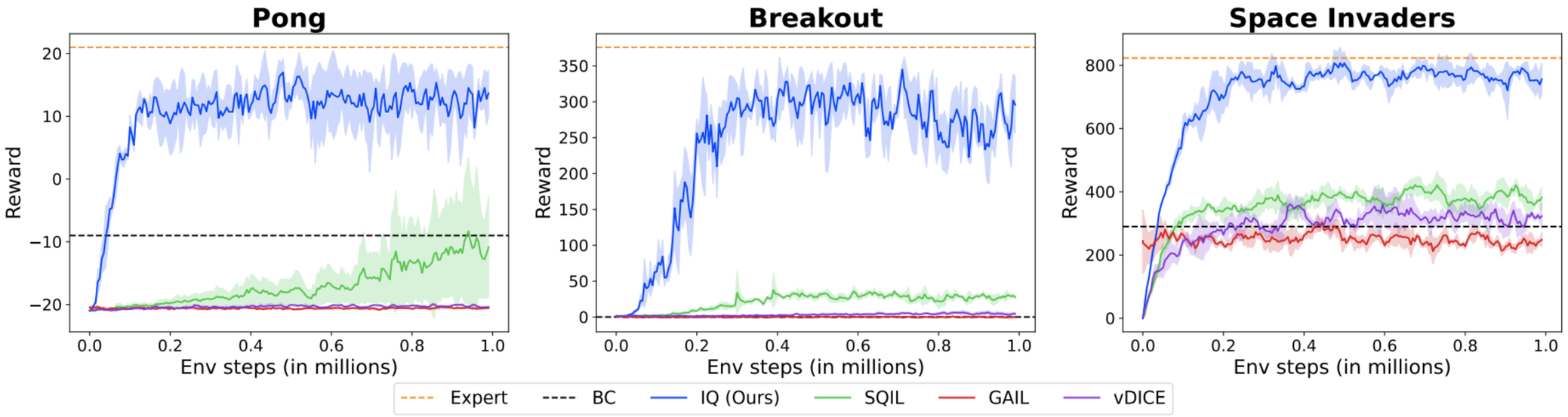
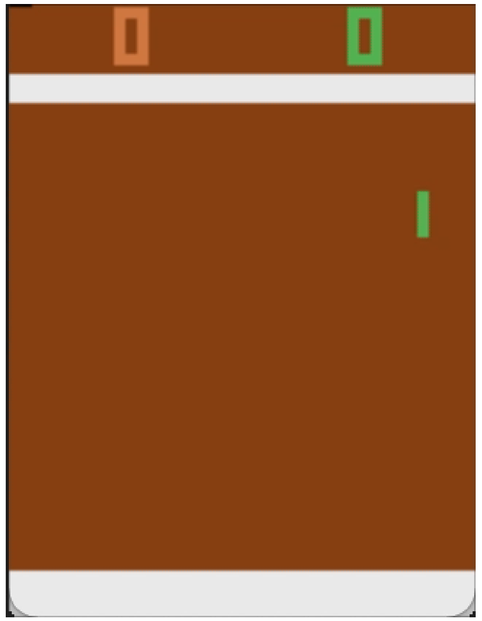
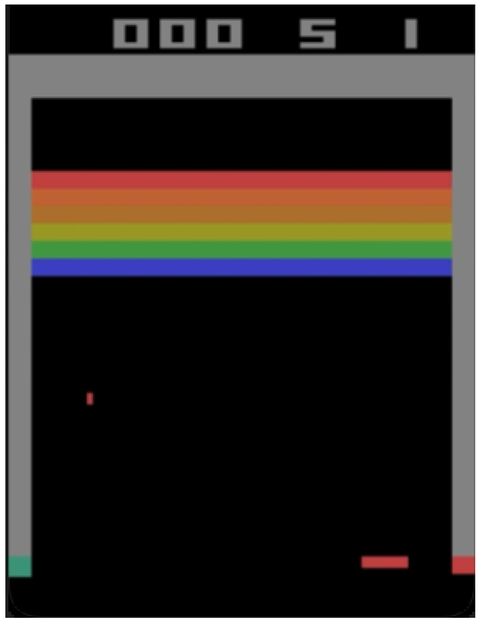


Figure 3: **Atari Results**. We show the returns vs the number of env steps. (Averaged over 5 seeds)

Pong

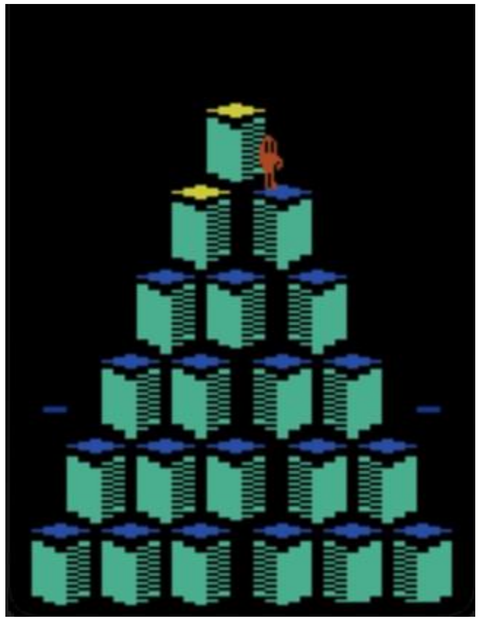


Breakout



Space Invaders

Q*bert



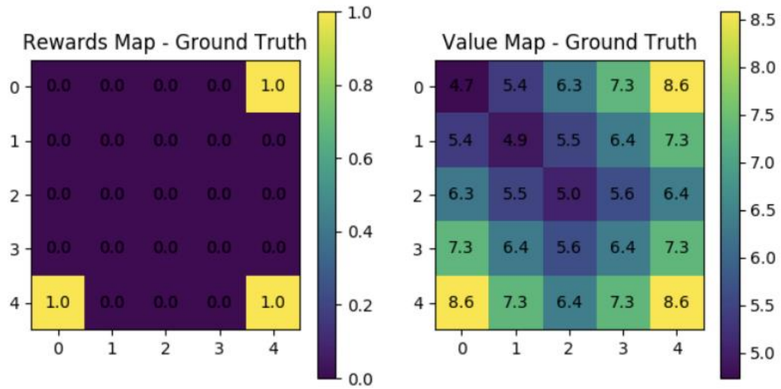
Results: Mujoco

Table 3: **Mujoco Results.** We show our performance on MuJoCo control tasks using a single expert trajectory.

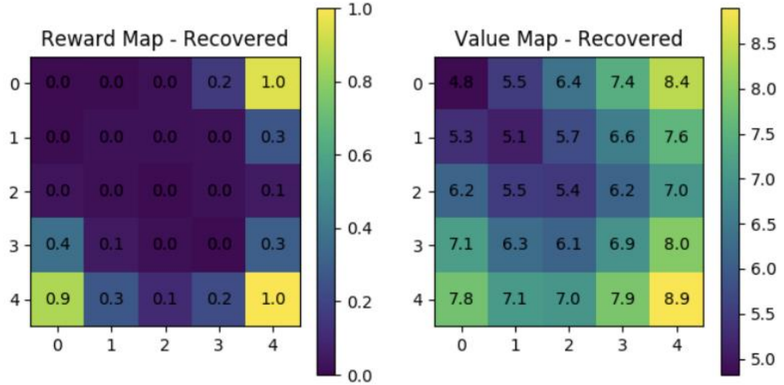
Task	GAIL	ValueDICE	IQ (Ours)	Expert
Hopper	3252.5	3312.1	**3546.4**	3532.7
Half-Cheetah	3080.0	3835.6	**5076.6**	5098.3
Walker	4013.7	3842.6	**5134.0**	5274.5
Ant	2299.1	1806.3	**4362.9**	4700.0

Recovered Rewards

Ground-truth



Recovered



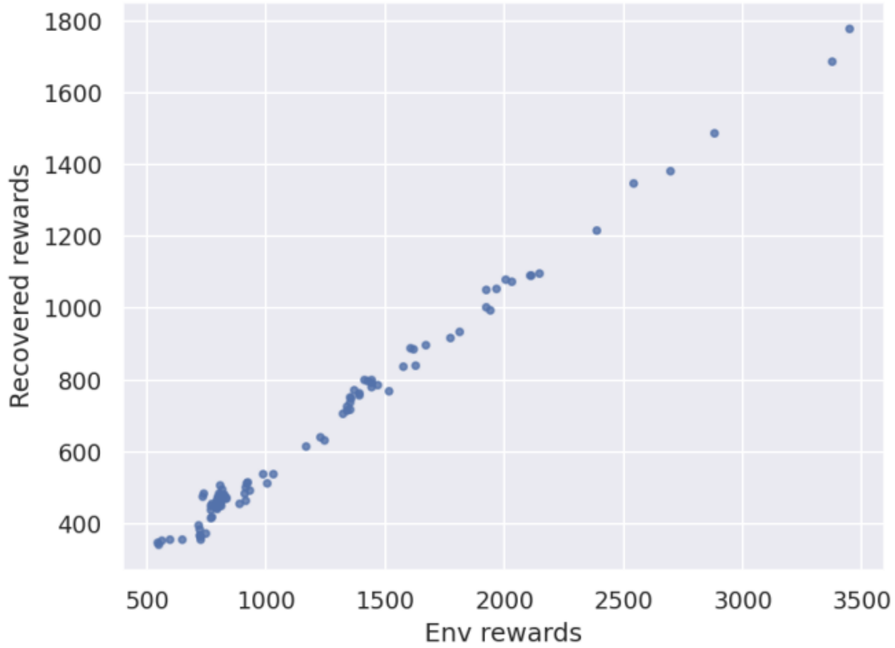


Figure 12: **Hopper correlations**

Figure 4: **Reward Visualization.** We use a discrete GridWorld environment with 5 possible actions: up, down, left, right, stay. Agent starts in a random state. (With 30 expert demos)

Extensions

1. Learning state-only reward functions
2. Imitation learning with only observations (ILO)

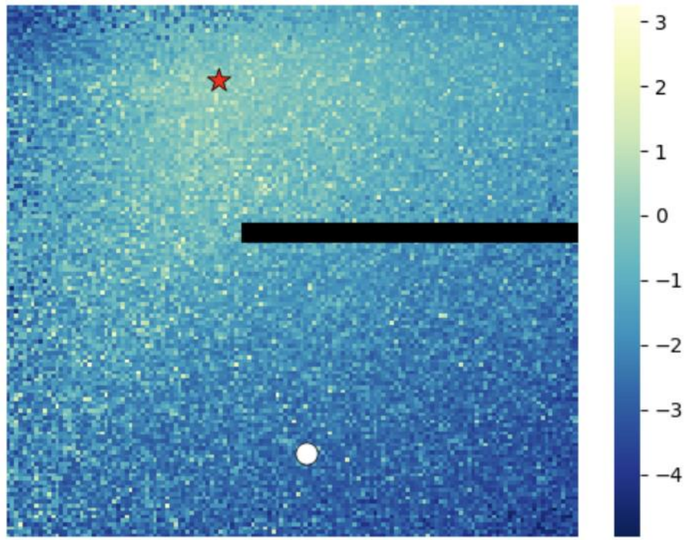


Figure 5: **State Rewards Visualization.** We visualize the state-only rewards recovered on a continuous control point maze task. The agent (white circle) has to reach the goal (red star) avoiding the barrier on right.

Table 6: **Results on ILO**. We show evironment returns using 1 and 10 expert demonstrations.

Env	1 demo	10 demos
CartPole	452 ± 50	485 ± 25
LunarLander	20 ± 102	220 ± 69
Hopper	2507 ± 345	3465 ± 51

Thank you

Paper

Code