IQ-Learn: Inverse soft-Q Learning for Imitation

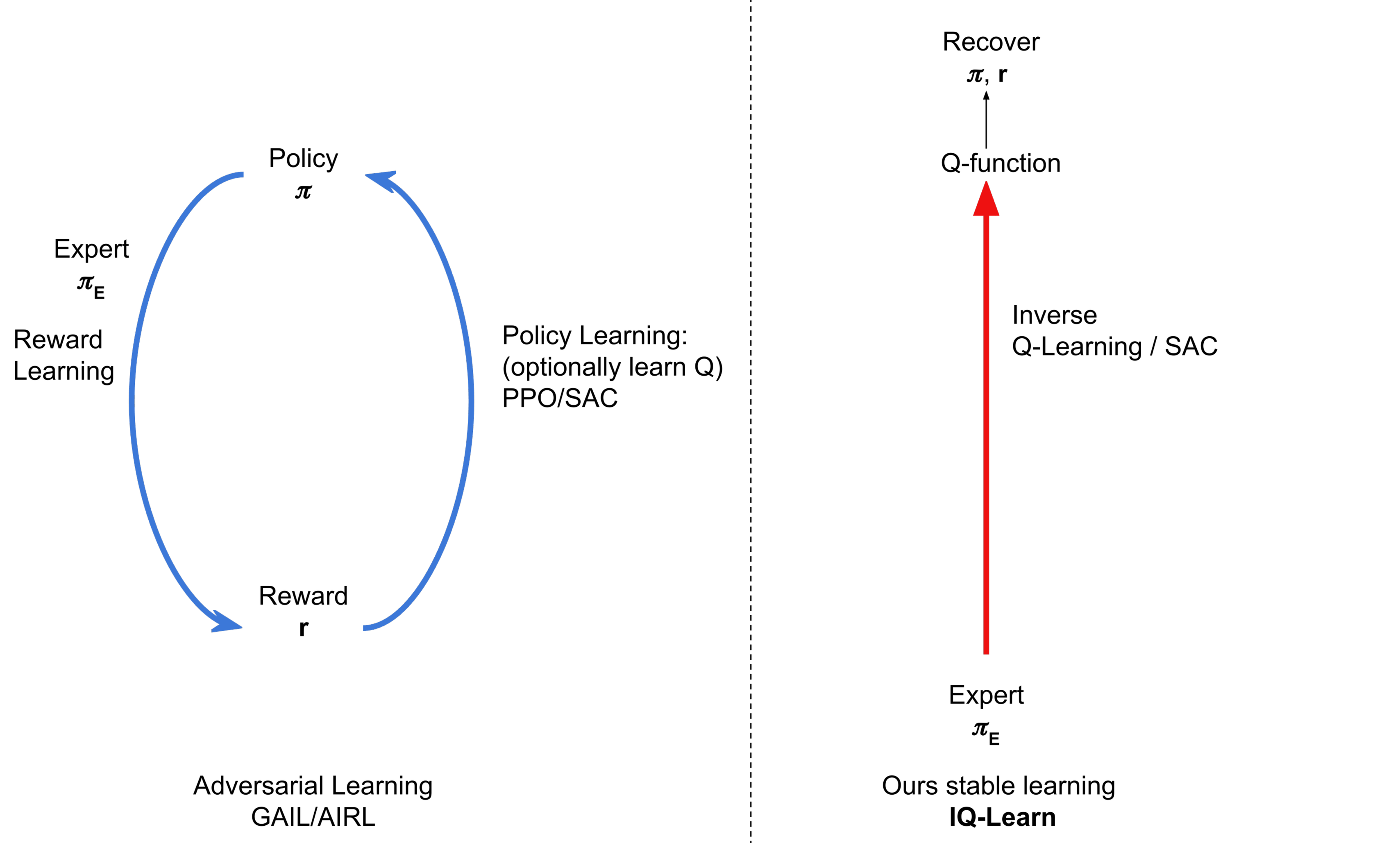
Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, Stefano Ermon

NEURAL INFORMATION PROCESSING SYSTEMS

SCAN ME

Inverse Q-Learning

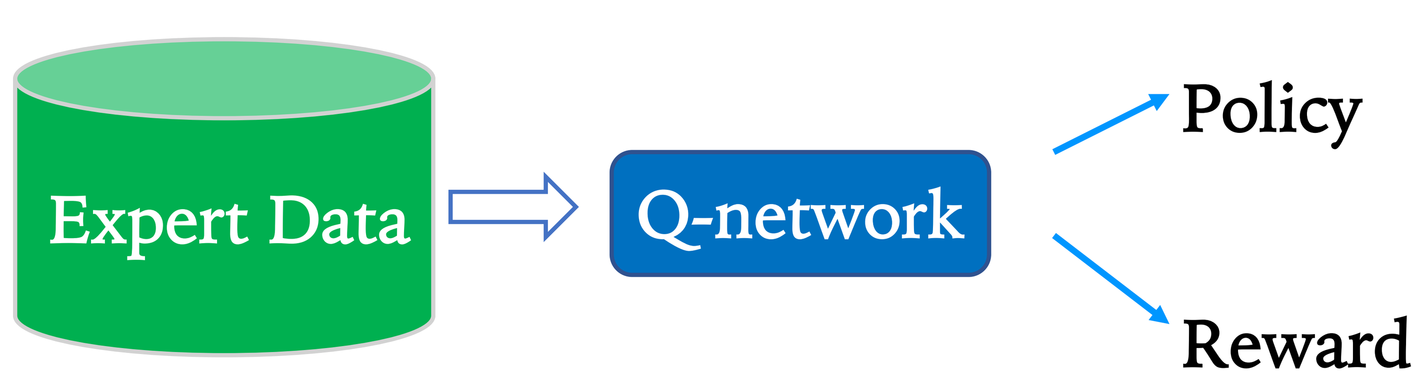
Learn Q-values from expert demos to recover both **optimal policy** and **rewards**



Adversarial Inverse RL	IQ-Learn
❌ Doesn't scale to complex envs	✅ Scales well to complex envs
❌ Difficult to convergence	✅ Convergence Guarantees
❌ Sensitive to hyperparameters	✅ Stable Simple Optimization
	✅ Works offline and online

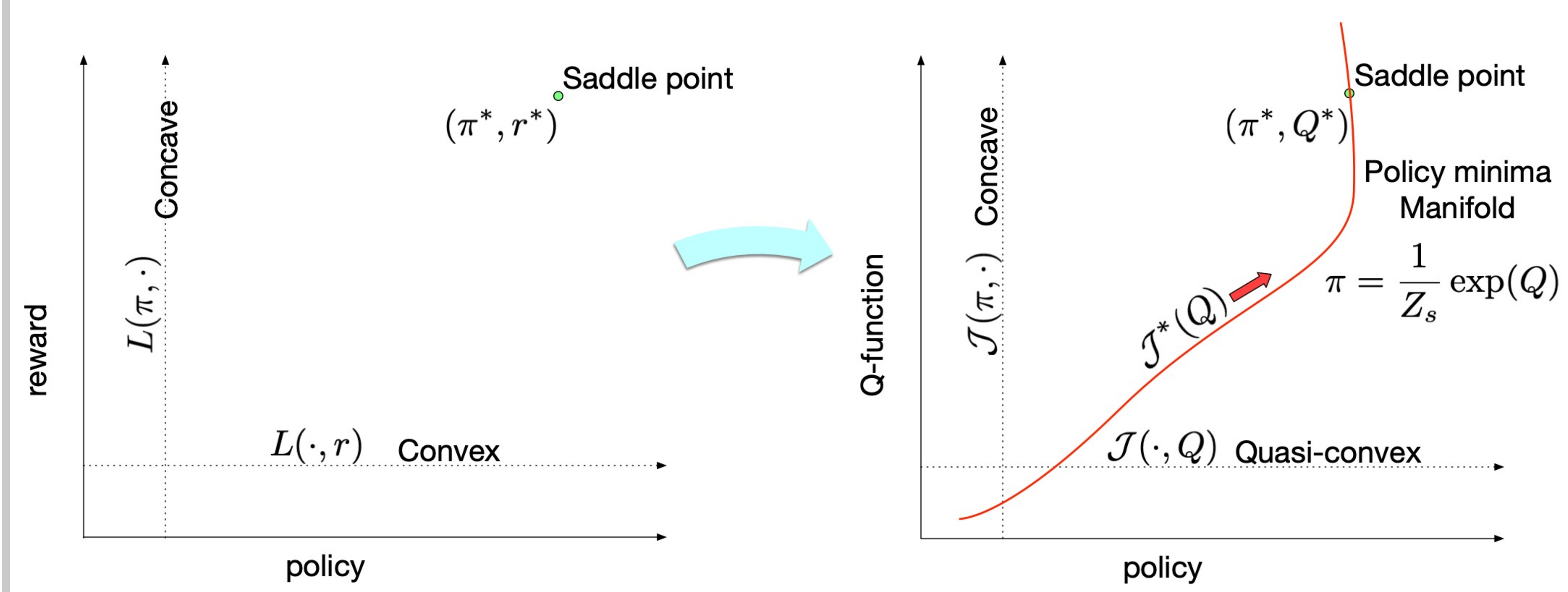
Method

IQ-Learn Algorithm: Modified Critic Update Rule



Can be implemented in **fewer than 15 lines** of code !!

Approach



Theorum:
Inverse RL ⇔ Inverse Q-Learning

Results

State-of-the-art in offline and online Imitation Learning

Offline IL

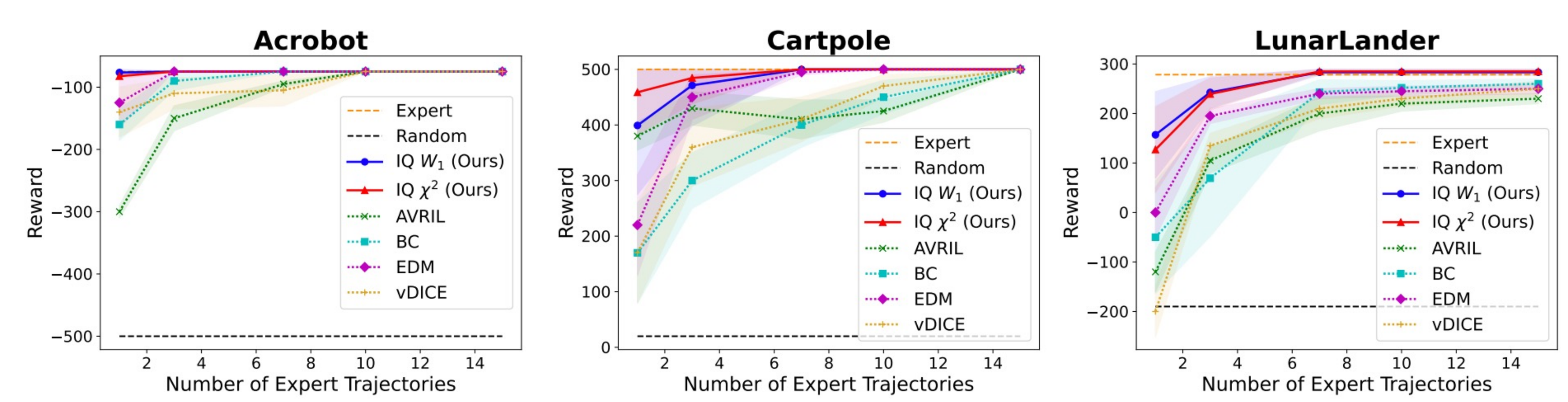


Figure 2: **Offline IL results.** We plot the average environment returns vs the number of expert trajectories.

Online IL

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

Atari Results

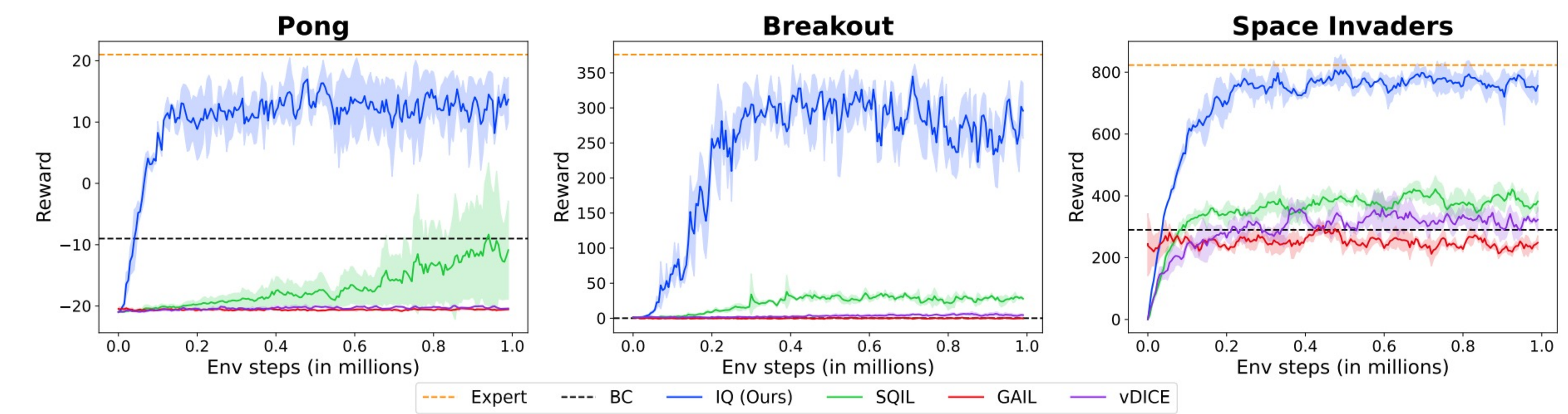


Figure 3: **Atari Results**. We show the returns vs the number of env steps. (Averaged over 5 seeds)

Outperform prior IL methods by **more than 3x**

Recovering Rewards

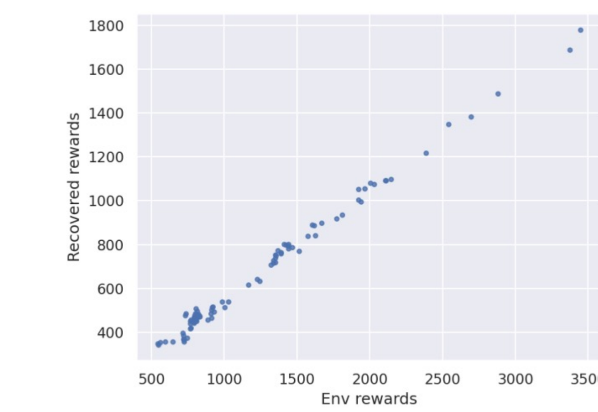


Figure 12: **Hopper correlations**

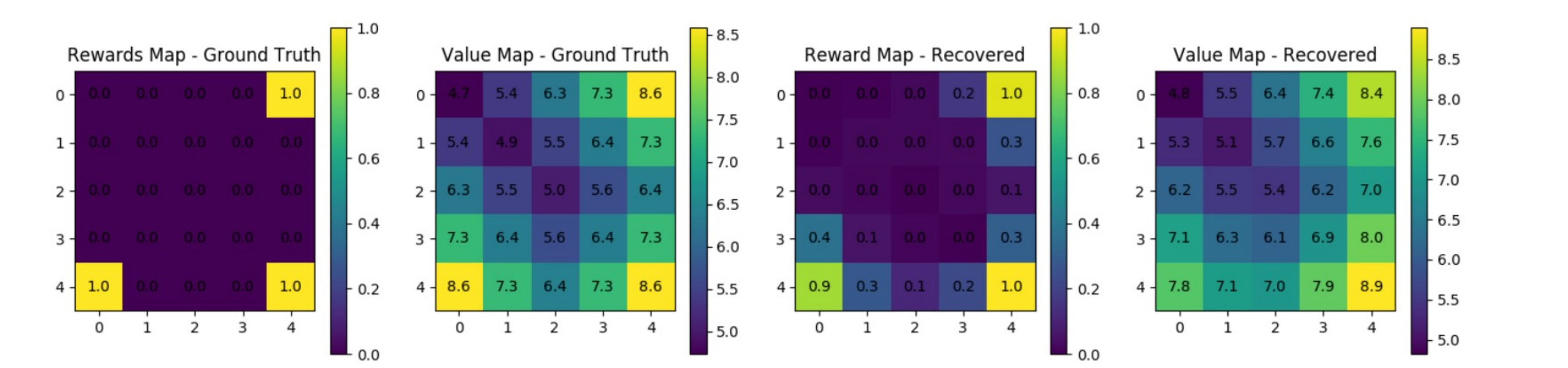


Figure 4: **Reward Visualization.** We use a discrete GridWorld environment with 5 possible actions: up, down, left, right, stay. Agent starts in a random state. (With 30 expert demos)

Imitation with Observations

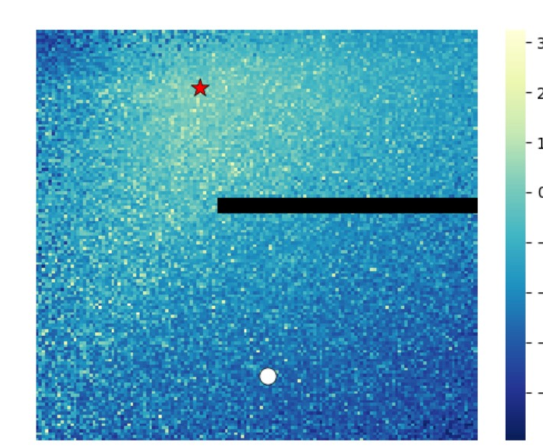


Figure 5: **State Rewards Visualization.** We visualize the state-only rewards recovered on a continuous control point maze task. The agent (white circle) has to reach the goal (red star) avoiding the barrier on right.

Table 8: **Results on ILO.** We show environment 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