Scalable Causal Imitation Learning

Eylam Tagor **Mingxuan Li** **Elias Bareinboim**

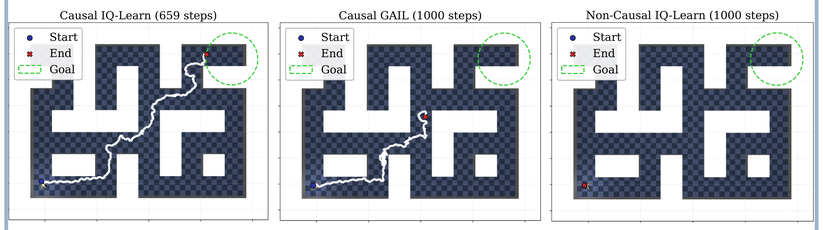
Causal Artificial Intelligence Lab · Department of Computer Science · Columbia University

Full paper

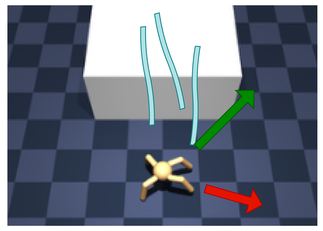
Reinforcement Learning Conference

Introduction & Motivation

- Imitation learning assumes the imitator observes everything the expert does ("No Unobserved Confounders"), though this is often not the case (expert-only sensors, shift of latent conditions such as wind, friction and payload).
- Under confounding, standard IL risks fitting spurious correlations from the demonstrations that hinder performance during deployment.
- Causal IL has emerged to negate this, but has remained restricted to low dimensionality and short horizon tasks.
- **We need imitation learning that is both scalable and causally informed.**



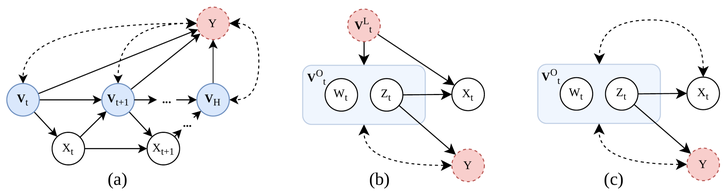
Example: Confounded AntMaze



Wind (blue) distorts compass (red) from true heading (green).

- The expert observes its true heading, but the imitator sees only a wind-deflected compass that correlates with the expert's turns in the demonstrations, because the latent wind drives both.
- While the compass seems to provide crucial (if inconsistent) information about the expert's decision-making process, it does not influence the expert policy at all.
- Learning to condition on the compass risks diverging from the expert policy and introduces susceptibility to distribution shift.

Causal Structure



- Conditioning on every observed variable opens a spurious path.
- The sequential π-backdoor criterion returns an adjustment set Z_t that excludes the compass, thus indifferent to the shift in W.

Why prior causal IL doesn't scale

Causal BC: inherits behavioral cloning's compounding error.

Causal GAIL: on-policy rollouts collect too few complete traversals for discriminator.

We need: off-policy IL that (a) treats demonstrations as a static buffer, (b) learns from self-collected transitions, and (c) propagates reward across long horizons.

Causal SQIL and Causal IQ-Learn

Inverse soft-Q objectives with causally adjusted states, using (z_t, x_t) instead of (s_t, a_t).

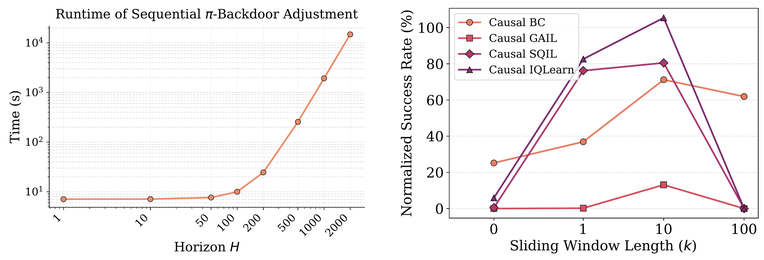
Causal SQIL: reward 1 on expert transitions, 0 on policy transitions; SAC on the combined replay buffer.

Causal IQ-Learn: learns a Q-function whose implicit reward is consistent with expert behavior.

Because Z_t satisfies the π-backdoor criterion the critic's value estimates are unconfounded, while TD learning carries the expert signal across the full horizon.

The adjustment layer is algorithm-agnostic!

Windowed π-backdoor adjustment



Sequential π-backdoor adjustment is **infeasible** at long horizons. We can shrink the problem using properties of continuous control tasks (**Assumption 1**):

- **Bounded influence.** Temporally localized physics and external forces.
- **Time-homogeneity.** Causal structural functions are identical across timesteps.

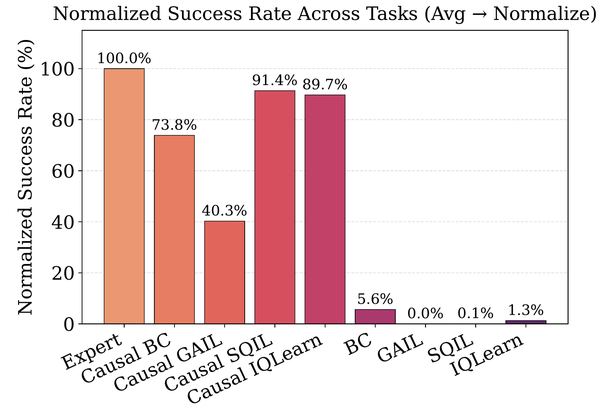
In Confounded AntMaze:

- A gust pushes the torso and compass, but the effect dissipates after a few steps.
- Maze and joint dynamics obey the same laws of physics every timestep.

Windowed adjustment. Solve the sequential π-backdoor once on a proxy graph of horizon $k+1$, then transfer the pattern to the rest of the timesteps. Under bounded influence and time-homogeneity, the windowed adjustment sets $\mathbf{Z}_t$ satisfy the sequential π-backdoor criterion (**Theorem 1**). Adjustment cost drops from $O(H)$ to $O(k)$.

Bounded influence holds empirically: best k values are between 1 and 10. 0 (true Markov) loses information, high values like 100 bloat the representation.

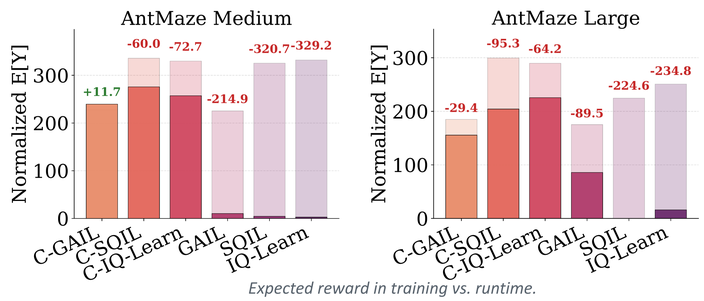
Experimental Results



Success rate (%) by task

Algorithm	Ant-Med	Ant-Lg	Hum-Med	Hum-Lg
Expert	87.6	55.9	33.8	7.0
Causal BC	77.9	39.8	10.4	8.0
Causal GAIL	74.1	7.3	0.0	0.0
Causal SQIL (ours)	**90.7**	**45.0**	**24.7**	**8.0**
Causal IQ-Learn (ours)	**84.3**	**58.9**	**19.1**	**3.0**
BC	0.0	0.0	5.4	5.0
GAIL	0.0	0.0	0.0	0.0
SQIL	0.0	0.0	0.1	0.0
IQ-Learn	0.0	0.0	2.4	0.0

Confounding is invisible during training



Expected reward in training vs. runtime.

During training, causal and non-causal imitators are comparable. The gap emerges only at runtime due to unobserved confounders causing distributional changes from the expert demonstrations, affecting only the non-causal methods. **Confounding bias cannot be revealed by in-distribution evaluation.**

Acknowledgements

This research is supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

Code: github.com/CausalAILab/Scalable-Causal-Imitation-Learning