DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF OXFORD

Causal Imitation Learning under Expert-Observable and Expert-Unobservable Confounding

Daqian Shao, Thomas Kleine Buening, Marta Kwiatkowska

ICLR 2026

Imitation Learning (IL)

- Imitation learning aims to learn a policy that mimics the behaviour of an expert by learning from its demonstrations.
- However, it has been observed in practice that IL algorithms produce suboptimal and unsafe policies, inconsistent with classical IL theory that states with infinite data, the imitator should be value-equivalent to the expert.
- One of the key reasons for this is in real datasets, spurious correlation and hidden variables exist, which causal the assumptions in classical IL theory to break. Therefore, new theory and methodology are required.

Background: Instrumental Variables

DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF OXFORD

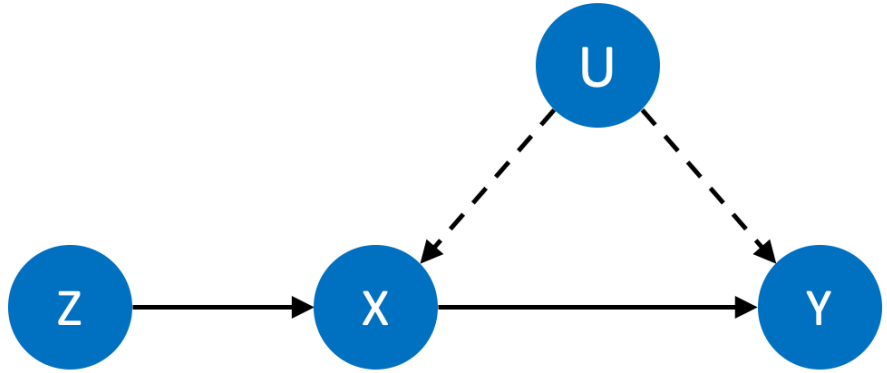
Consider a causal model that specifies some outcome $\boldsymbol{Y}$ given treatment $\boldsymbol{X}$ with a hidden confounder $\boldsymbol{U}$:

$$Y = f(X) + \varepsilon(U) \quad with \quad \mathbb{E}[\varepsilon(U)] = 0 \quad \text{but} \quad \mathbb{E}[\,\varepsilon(U) \mid X\,] \neq 0.$$

Standard regressions (e.g., OLS) generally **fail** to estimate the **causal relationship** between X and Y, i.e., $f(X)$, since:

$$\mathbb{E}[\,Y \mid X\,] = f(X) + \mathbb{E}[\,\varepsilon(U) \mid X\,] \neq f(X)$$

Instrumental variables (IVs) are random variables independent to the hidden confounder and only affect the outcome through the action.



With **IVs**, it becomes possible to identify $f(X)$. The problem can be reduced to solving a set of **Conditional Moment Restrictions (CMRs):**

$$\mathbb{E}[\,Y - f(X) \mid Z\,] = 0$$

This is not trivial since it is an inverse problem!

Spurious Correlations in IL

DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF OXFORD

Expert	Spurious Correlations to the Imitator	Hidden Confounder
Flying a drone	Was the observed demonstrations intentional by the expert or influenced by the wind?	Wind
	Did the expert change direction to avoid a no-fly zone or due to other observations?	Map of no-fly zone
Driving a car	Did the car slow down because we are going uphill or something else?	small hills
	Did the car slow down because the breaking light is on or something else?	The safety hazard
Pricing plane tickets	Was the increase in price caused by the fluctuating fuel price or the observable data?	Fuel price
	Was the increase in price caused by unknown expert knowledge or the observable data?	Expert-exclusive knowledge

How to unify and categorise these hidden confounders?

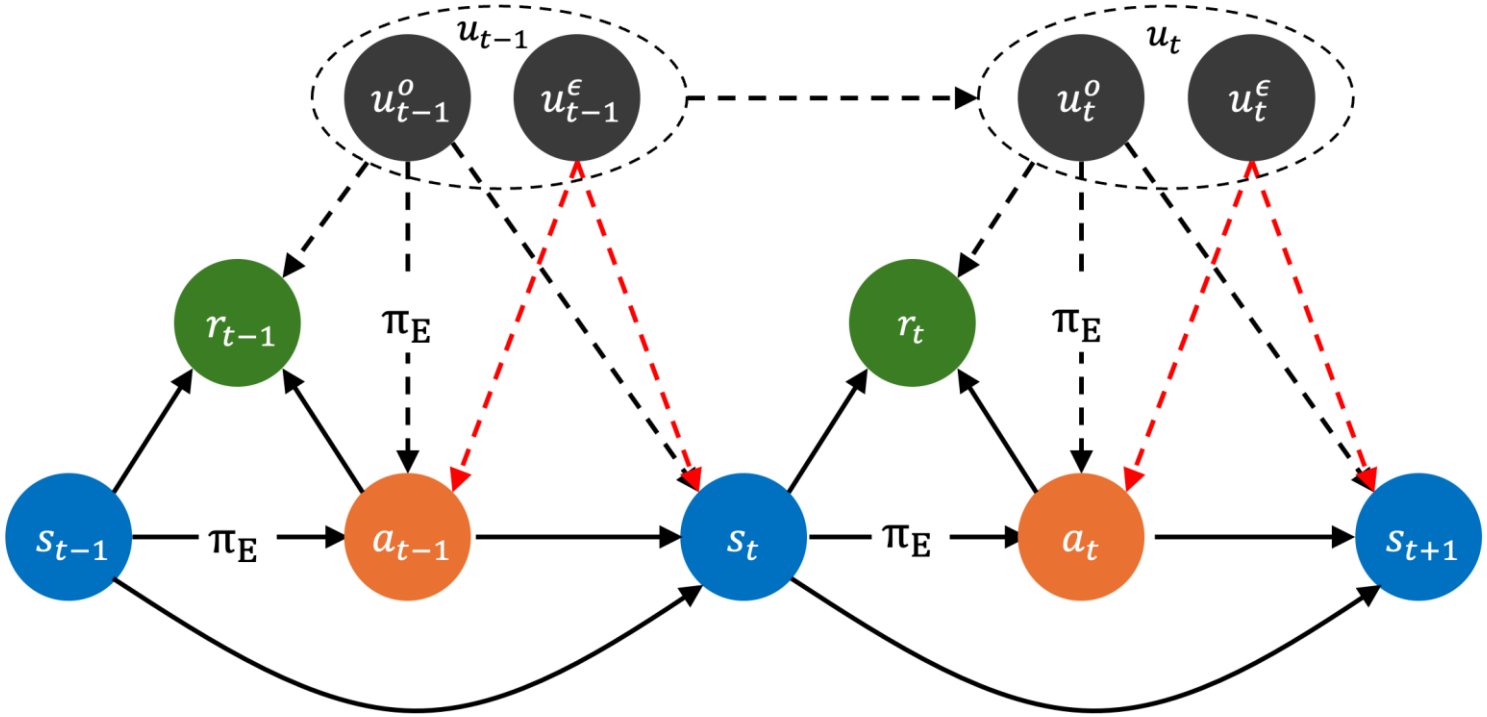
Some are known to or observed by the expert!

Others are hidden to both the expert and imitator.

A General Framework of Causal IL

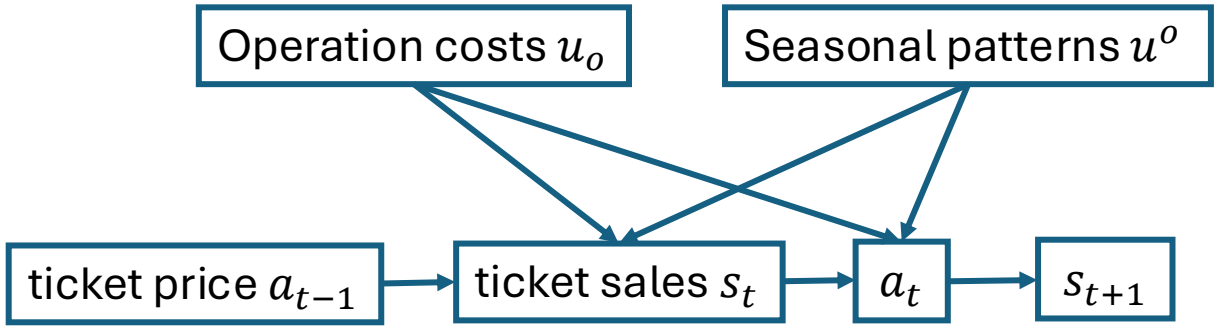
MDPs with hidden confounders $(\mathcal{S}, \mathcal{A}, \mathcal{U}, P, r, \mu_0, T)$.
Crucially, we model hidden confounder as **two parts: $u_t = (u_t^o, u_t^\varepsilon)$.**

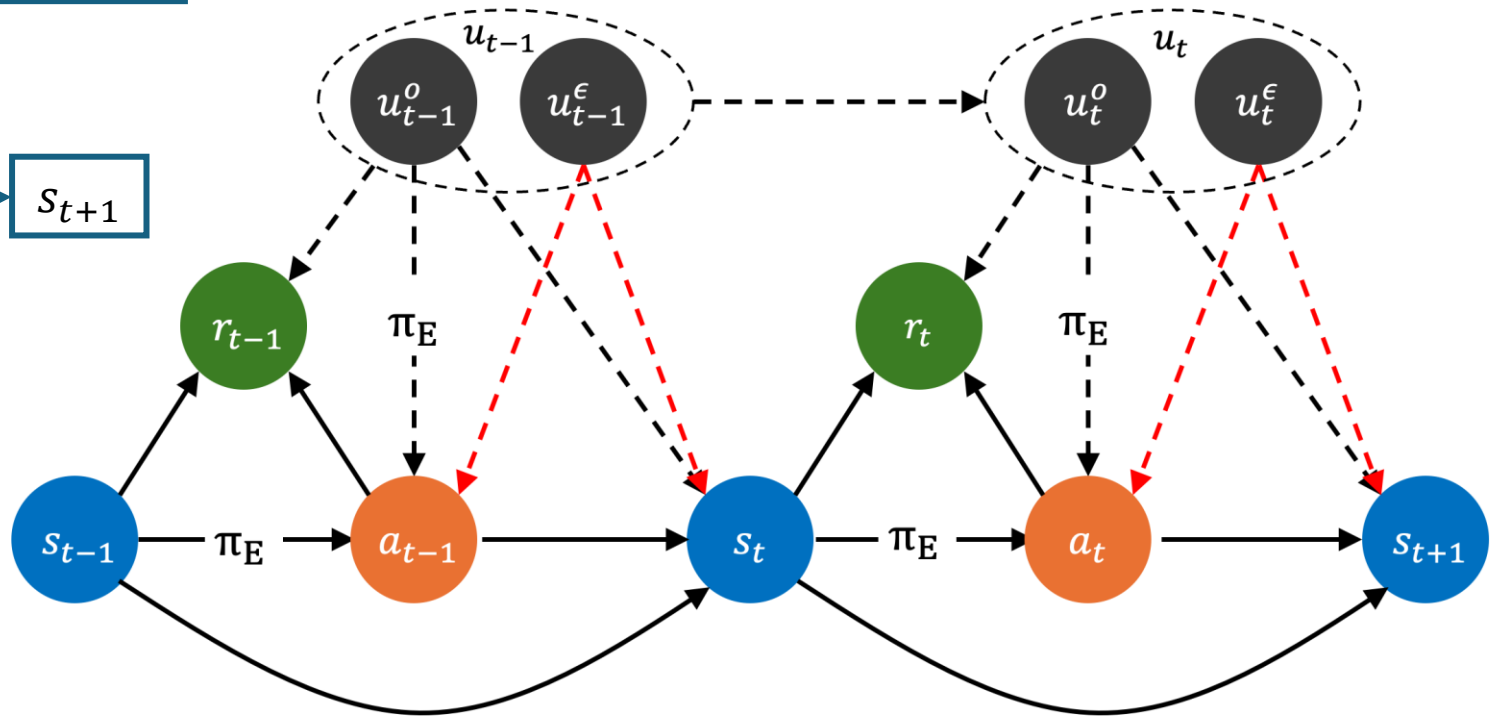
- u_t^ε is **unobservable to both the expert and imitator** (**confounding noise**)
  - i.e., wind, small hills and fuel price
- u_t^o are **additional information** that **only the expert observes.**
  - i.e., no-fly zone, safety hazard and expert-exclusive knowledge



A General Framework of Causal IL

MDPs with hidden confounders $(\mathcal{S}, \mathcal{A}, \mathcal{U}, P, r, \mu_0, T)$.
Crucially, we model hidden confounder as **two parts: $u_t = (u_t^o, u_t^\epsilon)$.**





A General Framework of Causal IL

DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF OXFORD

We want to learn a history-dependent imitator policy π_h from confounded expert demonstrations following expert policy π_E in this MDP with hidden confounders.

With hidden confounders, we need to infer the expert's **true intention**
This is a **causal inference problem**, and for identifiability, we need the following assumptions:

Assumption 3.2 (Confounding Noise Horizon). For every t, the confounding noise u_t^ε has a horizon of k where $1 \leq k < T$. More formally, $u_t^\varepsilon \perp\!\!\!\perp u_{t-k}^\varepsilon \; \forall t > k$.

This assumption **holds** when:

- The effect of the u_t^ε **diminishes over time**: wind, hills, environment noises.
- u_t^ε **becomes observable** at a future time: Fuel price, exclusive information

A General Framework of Causal IL

DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF OXFORD

Assumption 3.3 (Additive Noise)**.** The structural equation that generates the actions in the observed trajectories is

$$a_t = \pi_E(s_t, u_t^o) + u_t^\varepsilon, \tag{2}$$

where w.l.o.g. $\mathbb{E}[u_t^\varepsilon] = 0$ as any non-zero expectation of u_t^ε can be included as a constant in π_E.

This is a **standard assumption** in causal inference and IV frameworks, which is required for the causal effect to be identifiable。

Causal IL as CMRs

The typical target for IL would be $\pi_E(s_t, u_t^o)$. However, since **we don't observe u_t^o**, the best we can do is to estimate its **expectation conditional** on the observable **trajectory history h_t**, which minimises least squares:

$$\pi_h(h_t) \coloneqq \mathbb{E}_{\mathbb{P}(u_t^o|h_t)}[\pi_E(s_t, u_t^o)] = \mathbb{E}[\pi_E(s_t, u_t^o) \mid h_t]$$

With the confounding noise horizon k, we can use the **trajectory history h_{t-k}** k-steps ago as the **instrument** to break the spurious correlation:

$$\begin{aligned}
\mathbb{E}[a_t \mid h_{t-k}] &= \mathbb{E}\left[\mathbb{E}[a_t \mid h_t] \mid h_{t-k}\right] \\
&= \mathbb{E}[\pi_h(h_t) \mid h_{t-k}] + \mathbb{E}[\mathbb{E}[u_t^\varepsilon \mid h_t] \mid h_{t-k}] \\
&= \mathbb{E}[\pi_h(h_t) \mid h_{t-k}] + \mathbb{E}[u_t^\varepsilon \mid h_{t-k}] \\
&= \mathbb{E}[\pi_h(h_t) \mid h_{t-k}] + \mathbb{E}[u_t^\varepsilon] \\
&= \mathbb{E}[\pi_h(h_t) \mid h_{t-k}].
\end{aligned}$$

- Definition of π_h
- Law of total expectation
- u_t^ε and u_{t-k}^ε independent
- $\mathbb{E}[u_t^\varepsilon]$ zero expectation

- we can check that the **h_{t-k}** indeed satisfy the conditions of IV

This is a **CMRs problem**! $\mathbb{E}[a_t - \pi_h(h_t) \mid h_{t-k}] = 0$

Causal IL as CMR

We can optimize the following **CMR error** by adopting IV regression algorithms

$$\sqrt{\mathbb{E}\big[\mathbb{E}[a_t - \hat{\pi}_h(h_t)|h_{t-k}]^2\big]} = \|\mathbb{E}[a_t - \hat{\pi}_h(h_t)|h_{t-k}]\|_2.$$

We propose the following algorithm based on the IV regression algorithm DML-IV.

Algorithm 1 DML-IL

input Dataset $\mathcal{D}_E$ of expert demonstrations, confounding noise horizon k
Initialize the roll-out model $\hat{M}$ as a Gaussian mixture model
repeat
    Sample (h_t, a_t) from data $\mathcal{D}_E$
    Fit the roll-out model $(h_t, a_t) \sim \hat{M}(h_{t-k})$ to maximize the log likelihood
until convergence
Initialize the expert model $\hat{\pi}_h$ as a neural network
repeat
    Sample h_{t-k} from $\mathcal{D}_E$
    Generate $\hat{h}_t$ and $\hat{a}_t$ using the roll-out model $\hat{M}$
    Update $\hat{\pi}_h$ to minimise the loss $\ell := \|\hat{a}_t - \hat{\pi}_h(\hat{h}_t)\|_2$
until convergence
return A history-dependent imitator policy $\hat{\pi}_h$

Theoretical Guarantees

To **bound the imitation gap** (performance gap to expert), we need to control the following:

1. **Information about the hidden confounders** that can be inferred from trajectory histories;
   - **Total Variation (TV) distance** between the distribution of $\boldsymbol{u_o^t}$ **and** $\mathbb{E}[\boldsymbol{u_o^t} \mid \boldsymbol{h_t}]$ along π_E

2. **ill-posedness** (or identifiability) of the set of CMRs (measures the **strength** of the instrument);
   - **ill-posedness** $\boldsymbol{\nu(\Pi, k)}$ (Dikkala et al., 2020): a function of policy class Π and confounding horizon k.

3. **Disturbance of the confounding noise** to the states and actions at test time.
   - **c-TV stability**, which bounds TV induced by noise when added to state/actions.

Theorem 4.5 (Imitation Gap Bound). *Let $\hat{\pi}_h$ be the learnt policy with CMR error ε and let $\nu(\Pi, k)$ be the ill-posedness of the problem. Assume that $\delta_{TV}(u_t^o, \mathbb{E}_{\pi_E}[u_t^o|h_t]) \leq \delta$ for $\delta \in \mathbb{R}^+$, $P(u_t^\varepsilon)$ is c-TV stable and π_E is deterministic. Then, the imitation gap is upper bounded by*

$$J(\pi_E) - J(\hat{\pi}_h) \leq T^2\big(c\varepsilon\nu(\Pi, k) + 2\delta\big) = \mathcal{O}\big(T^2(\delta + \varepsilon)\big).$$

Theoretical Guarantees

In addition, we show that the upper bounds on the imitation gap derived by prior works[4][5] are **special cases** of Theorem 4.5.

Corollary 4.6. *In the special case that $u_t^o = 0$, i.e., there are no expert-observable confounders, or $u_t^o = \mathbb{E}_{\pi_E}[u_t^o|h_t]$, i.e., u_t^o is $\sigma(h_t)$ measurable (all information about u_t^o is contained in the history), the imitation gap is upper bounded by*

$$J(\pi_E) - J(\hat{\pi}_h) \leq T^2\big(c\varepsilon\nu(\Pi, k)\big) = \mathcal{O}\big(T^2\varepsilon\big),$$

which coincides with Theorem 5.1 of Swamy et al. (2022a).

Corollary 4.7. *In the special case that $u_t^\varepsilon = 0$, if the learnt policy has optimisation error ε, the imitation gap is upper bounded by*

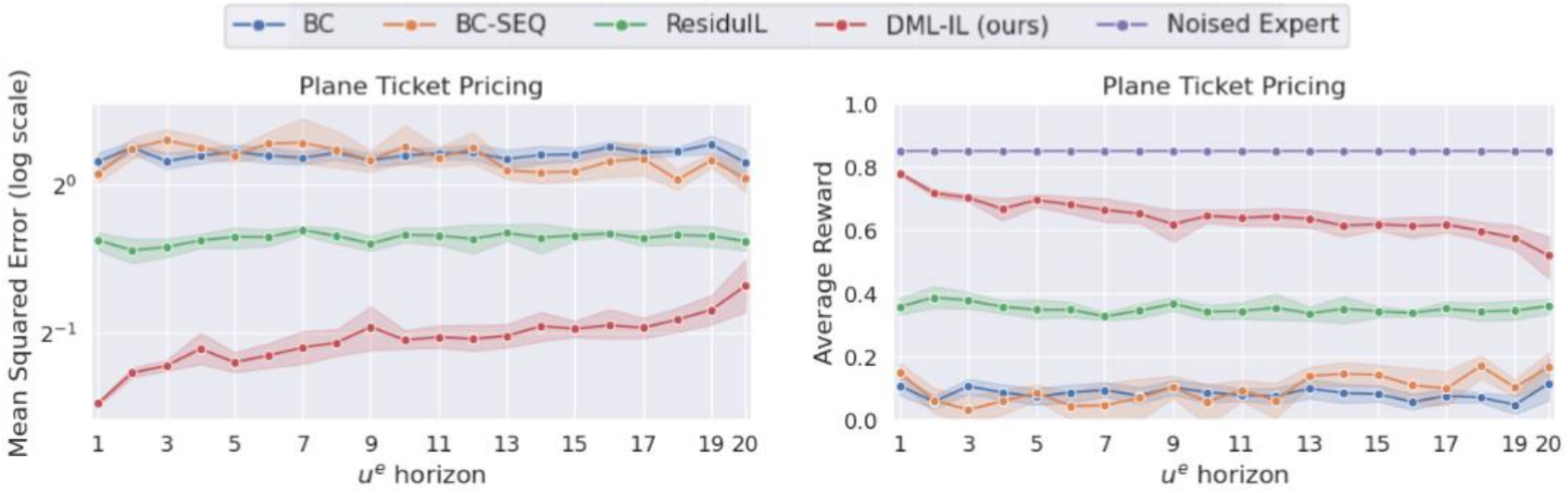
$$J(\pi_E) - J(\hat{\pi}_h) \leq T^2\left(\frac{2}{\sqrt{\dim(A)}}\varepsilon + 2\delta\right),$$

which is a concrete bound that extends the abstract bound in Theorem 5.4 of Swamy et al. (2022b).

[4] Gokul Swamy et al,. Causal imitation learning under temporally correlated noise, ICML 2022
[5] Gokul Swamy et al,. Sequence model imitation learning with unobserved context, Neurips 2022

Experimental Results

Results for a **Toy example** of airline ticket pricing environment

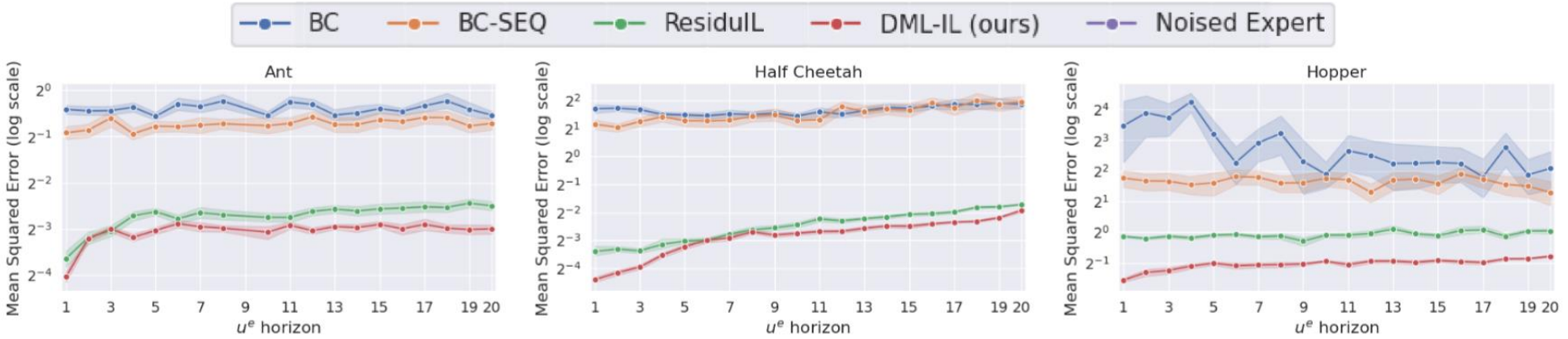


(a) MSE in log scale, lower is better. (b) Average reward, higher is better.

Figure 4: The MSE and the average reward in the airline ticket environment.

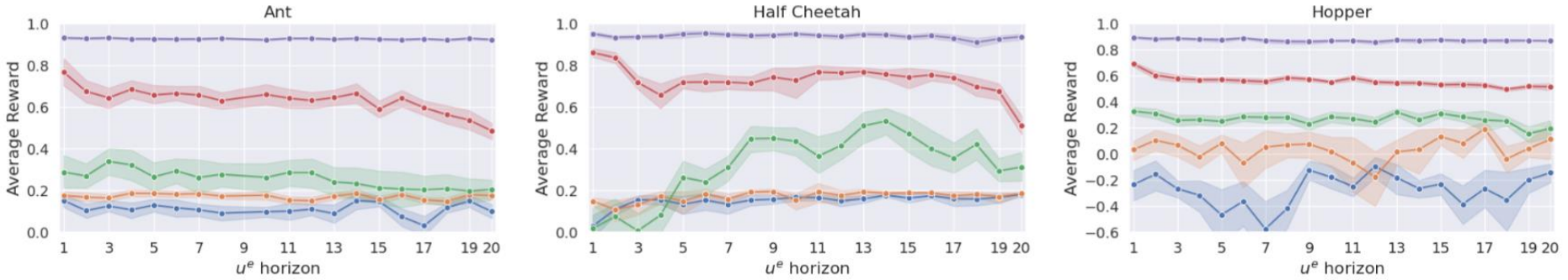
We also showed theoretically that the **ill-posedness** of the problem **monotonically increases** as the confounding horizon increases.

Experimental Results



(a) Ant (b) Half Cheetah (c) Hopper

Figure 2: MSE in the three Mujoco environments. Lower values are better.



(a) Ant (b) Half Cheetah (c) Hopper

Figure 3: The average reward in Mujoco environments. Higher values are better.

Results for three **Mujoco** environments

Conclusion

- We proposed a general framework for confounded IL, where the hidden confounders are partially observable to the expert
- IL in this setting can be reduced to a CMRs problem and we proposed algorithm DML-IL to solve these CMRs and imitate the expert
- Theoretical upper bound on the imitation gap for DML-IL
- Strong empirical performance against other causal IL algorithms

- Main Limitation: We assume there exist a confounding horizon k, and more importantly, we assume knowledge of the confounding noise horizon k or an upper bound on it for DML-IL. Unfortunately, the value of k generally cannot be verified empirically. Therefore, we rely on a sensible choice of k by the user based on the environment.