A UNIFYING FRAMEWORK FOR CAUSAL IMITATION LEARNING WITH HIDDEN CONFOUNDERS

Daqian Shao
University of Oxford
Oxford, UK
daqian.shao@cs.ox.ac.uk

Thomas Kleine Buening
The Alan Turing Institute
London, UK

Marta Kwiatkowska
University of Oxford
Oxford, UK

ABSTRACT

We propose a general and unifying framework for causal Imitation Learning (IL) with hidden confounders that subsumes several existing settings. Our framework accounts for two types of hidden confounders: (a) those observed by the expert but not the imitator, and (b) confounding noise hidden to both. By leveraging trajectory histories as instruments, we reformulate causal IL into Conditional Moment Restrictions (CMRs). We propose DML-IL, an algorithm that solves these CMRs via instrumental variable regression, and upper bound its imitation gap. Empirical evaluation on continuous state-action environments, including Mujoco tasks, shows that DML-IL outperforms state-of-the-art causal IL methods.

1 INTRODUCTION

Imitation Learning (IL) aims to learn a policy that replicates expert behaviour from demonstrations. While classical IL theory suggests that, with infinite data, the IL error should vanish (Ross et al., 2011), practical implementations often yield suboptimal and unsafe behaviours (Lecun et al., 2005, Kuefler et al., 2017, Bansal et al., 2018). Prior work attributes these failures to various factors, including spurious correlations (de Haan et al., 2019, Codevilla et al., 2019, Pfrommer et al., 2023), temporal noise (Swamy et al., 2022b), expert-exclusive knowledge (Choudhury et al., 2017, Chen et al., 2019, Swamy et al., 2022a, Vuorio et al., 2022), causal delusions (Ortega & Braun, 2008, Ortega et al., 2021), and covariate shifts (Spencer et al., 2021). However, these studies address individual factors in isolation, whereas in practice multiple challenges coexist, making partial solutions insufficient. This calls for a more holistic approach that accounts for multiple confounding factors simultaneously.

We propose a unifying framework for causal imitation learning that models hidden confounders—variables present in the environment but not recorded in demonstrations. Importantly, we distinguish between *expert-observable* confounders, which influence expert decisions but are not accessible to the imitator, and *expert-unobservable* confounders, which introduce spurious correlations and remain hidden from both the imitator and the expert. As a result, our framework generalises prior settings and enables a broader, more realistic problem formulation.

In this unifying framework, we propose an IL method that leverages trajectory histories as Instrumental Variables (IVs) to mitigate spurious correlations caused by expert-unobservable confounders. Additionally, by learning a history-dependent policy, we can infer information about expert-observable confounders, which enables us to better imitate the expert despite lacking access to said variables. We show that IL in our framework can be reformulated as set of Conditional Moment Restrictions (CMRs)—a well-studied problem in econometrics and causal inference, which allows us to design practical algorithms with theoretical guarantees on the imitation gap.

Main Contributions. In summary, our main contributions are as follows:

- A unifying framework for causal IL (Section 3) incorporating both expert-observable and expert-unobservable confounding variables to unify and generalise many of the settings in prior work.
- Reformulation of causal IL in our framework in terms of solving CMRs, by leveraging trajectory histories as instruments to learn a history-dependent policy (Section 4).

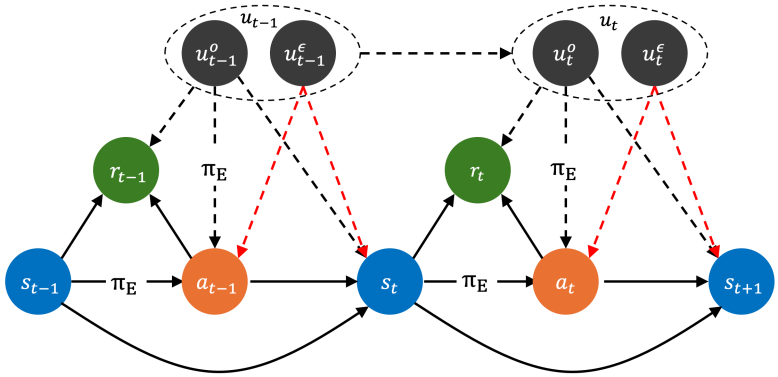


Figure 1: A causal graph of MDPs with hidden confounders $u_t = (u^o_t, u^\varepsilon_t)$. The black dotted lines represent the causal effect of the expert-observable confounder u^o_t, which directly affects a_t because the expert can observe u^o_t. It also directly affects s_{t+1} and r_t. The red dotted lines represent the causal effect of the expert-unobservable u^ε_t, which acts as confounding noise and directly affects the states and actions. u^ε_t does not directly affect r_t (following Swamy et al. (2022b)) because the expert policy does not take u^ε_t into account, and letting u^ε_t directly affect r_t would only add noise to the expected return.

- DML-IL, a novel algorithm for causal IL in our general framework, for which we prove an upper bound on the imitation gap that recovers prior works' results as special cases (Theorem 4.5).
- Empirical evaluation of our algorithm in challenging instances, where both types of confounders are present in the environment, demonstrating it outperforms SOTA methods (Section 5).

We discuss additional related work extensively in Appendix A and explain in detail how previously studied causal IL frameworks are generalised by our proposed unifying framework in Appendix B.

2 Instrumental Variables and Conditional Moment Restrictions

We first briefly introduce the concept of Instrumental Variables (IVs) and its connection to Conditional Moment Restrictions (CMRs). Consider a structural model for outcome Y and treatment X:

$$Y = f(X) + \varepsilon(U) \quad \text{with} \quad \mathbb{E}[\varepsilon(U)] = 0, \tag{1}$$

where U is a hidden confounder that affects both X and Y so that $\mathbb{E}[\varepsilon(U) \mid X] \neq 0$. Due to the presence of this hidden confounder, standard regressions (e.g., ordinary least squares) generally fail to produce consistent estimates of the causal relationship between X on Y, i.e., $f(X)$. If we only have observational data, a classic technique for learning f is IV regression (Newey & Powell, 2003). An IV Z is an observable variable that satisfies the following conditions:

- *Unconfounded Instrument*: $Z \perp\!\!\!\perp U$;
- *Relevance*: $\mathbb{P}(X|Z)$ is not constant in Z;
- *Exclusion*: Z does not directly affect Y: $Z \perp\!\!\!\perp Y \mid (X, U)$.

Using IVs, we are able to formulate the problem of learning f into a set of CMRs (Dikkala et al., 2020), where we aim to solve for f satisfying $\mathbb{E}[Y - f(X) \mid Z] = 0$. In our work, we show that trajectory histories can be used as instruments to learn the causal relationship between states and expert actions by transforming the problem of causal IL into CMRs (Section 4).

3 A Unifying Framework for Causal Imitation Learning

MDPs with Hidden Confounders. We introduce a novel unifying framework for causal IL based on Markov Decision Processes (MDPs) with hidden confounders: $(\mathcal{S}, \mathcal{A}, \mathcal{U}, \mathcal{P}, r, \mu_0, T)$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and $\mathcal{U}$ is the confounder space. Parts of the hidden confounders

u_t may be available to the expert due to imperfect environment logging and expert knowledge. We model this by segmenting the hidden confounder into two parts $u_t = (u_t^o, u_t^\varepsilon)$, where u_t^o are observable to the expert and u_t^ε are not. Intuitively, u_t^o are additional information that only the expert observes and u_t^ε behave as confounding noise in the environment that affects both the state and action.[1] The transition function $\mathcal{P}(\cdot \mid s, a, (u^o, u^\varepsilon))$ depends on both hidden confounders, while the reward $r(s, a, u^o)$ does not depend on the confounding noise u^ε as it only directly affects the state and actions. Finally, μ_0 is the initial state distribution and T is the horizon of the problem. A causal graph illustrating these relationships is provided in Figure 1 and a motivating example of a dynamic airline ticket pricing environment is provided below.

Example 3.1. Consider an airline ticket pricing scenario Wright (1928), where the goal is to learn a pricing policy by imitating actual airline pricing based on expert-set profit margins. Suppose that seasonal patterns and external events are known only to experts, but missing from the dataset, serving as expert-observable confounders u_t^o. Meanwhile, actual airline prices are confounded (additively) by fluctuating operating costs, which are unknown to the experts when they set the profit margin and unobserved in the dataset, making them confounding noise u_t^ε. We conduct experiments on a toy environment inspired by this in Appendix D.1 and show that IL algorithms that do *not* distinguish between u_t^o and u_t^ε fail to correctly imitate the expert.

Causal Imitation Learning. We assume that an expert is demonstrating a task following some expert policy π_E (which we will specify later) and we observe a set of $N \geq 1$ expert demonstrations $\mathcal{D}_E = \{d_1, d_2, ..., d_N\}$. Each demonstration is a state-action trajectory $(s_1, a_1, ..., s_T, a_T)$, where, at each time step, we observe the state s_t and the action a_t taken in the environment, and the trajectory follows the transition function $\mathcal{P}(\cdot \mid s_t, a_t, (u_t^o, u_t^\varepsilon))$. Denote $h_t = (s_1, a_1, ..., s_{t-1}, a_{t-1}, s_t) \in \mathcal{H}$ as the trajectory history at time t, where $\mathcal{H} \subseteq \bigcup_{i=0}^{T-1} (\mathcal{S} \times \mathcal{A})^i \times \mathcal{S}$ is the set of all possible trajectory histories at different time steps. Importantly, we do not observe the reward and the sequence of confounders (u_t^o, u_t^ε). Given the observed trajectories, our goal is to learn a history-dependent policy $\pi_h : \mathcal{H} \to \Delta(A)$. We assume that our policy class Π is convex and compact. The Q-function of a policy $\pi_h \in \Pi$ is defined as $Q_\pi(s_t, a_t, u_t^o) = \mathbb{E}_{\tau \sim \pi_h}[\sum_{t'=t}^{T} r(s_{t'}, a_{t'}, u_{t'}^o)]$ and the return of a policy is given by $J(\pi) = \mathbb{E}_{\tau \sim \pi_h}[\sum_{t=1}^{T} r(s_t, a_t, u_t^o)]$, where τ is the trajectory following π_h.

This nuanced distinction between u_t^o and u_t^ε is crucial for determining the appropriate method for IL, and we begin with a motivating example to illustrate the importance of considering $u_t = (u_t^o, u_t^\varepsilon)$.

In order to learn a policy π_h that matches the performance of π_E, we need to break the spurious correlation between states and expert actions by inferring what the expert would do if we intervened and placed them in state s_t when observing u_t^o. Unfortunately, the causal inference literature (Shpitser & Pearl, 2008) tells us that, without further assumptions, it is generally impossible to identify π_E. To determine the minimal assumptions that allow π_E to be identifiable, we first observe that u_t^ε can be correlated for all time steps t, making it impossible to distinguish between the intended actions of the expert and the confounding noise. However, in practice, the confounding noise at far-apart time steps is often independent. For example, the effect of the confounding noise u_t^ε at time t on future states and actions often diminishes over time, which is typically the case for random environment noise such as wind. In addition, when the confounding noise u_t^ε at time t becomes observable at a future time t', e.g., operating costs become observable later on as in Example 3.1, the unobservable confounding noise at times t and t' becomes independent. We formalise this intuition as a *confounding noise horizon* k:

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 is essential for decoupling the spurious correlation between the state and action pairs. We also assume that the confounding noise is additive to the action, which is standard in causal inference (Pearl, 2000, Shao et al., 2024). Without this assumption, the causal effect becomes unidentifiable (see, e.g., (Balke & Pearl, 1994)) and the best we can do is to upper/lower bound it.

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

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

[1] In our framework, we allow the actual actions taken in the environment to be affected by the noise. Noise that only perturbs data records can be considered as a special case of our framework.

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.

4 CAUSAL IL AS CMRS

In this section, we demonstrate that performing causal IL in our framework is possible using trajectory histories as instruments by reformulating the problem as CMRs.

The typical target for IL would be the expert policy π_E itself. However, since the expert has access to privileged information, namely u_t^o, which the imitator does not, the best thing an imitator can do is to learn a history-dependent policy π_h to match the expert behaviour. A natural choice is the conditional expectation of $\pi_E(s_t, u_t^o)$ on the history h_t:

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

because the conditional expectation minimises the least squares criterion (Hastie et al., 2001) and π_h is the best predictor of π_E given h_t. In π_h, the distribution $\mathbb{P}(u_t^o \mid h_t)$ captures the information about u_t^o that can be inferred from trajectory histories.

Remark 4.1. *Learning π_h is not trivial. Policies learnt naively using behaviour cloning (i.e., $\mathbb{E}[a_t \mid h_t]$) fail to match π_E. In view of Equation* (2)*, we have that*

$$\mathbb{E}[a_t \mid h_t] = \mathbb{E}[\pi_E(s_t, u_t^o) \mid h_t] + \mathbb{E}[u_t^\varepsilon \mid h_t] = \pi_h(h_t) + \mathbb{E}[u_t^\varepsilon \mid h_t],$$

where $\mathbb{E}[u_t^\varepsilon \mid h_t] \neq 0$ due to the spurious correlation between u_t^ε and the trajectory history h_t. As a result, $\mathbb{E}[a_t \mid h_t]$ becomes biased, which can lead to arbitrarily worse performance compared to π_E.

Derivation of CMRs. Leveraging the confounding horizon from Assumption 3.2, it becomes possible to break the spurious correlation using the independence of u_t^ε and u_{t-k}^ε. We propose to use the k-step trajectory history $h_{t-k} = (s_1, a_1, ..., s_{t-k})$ as an instrument for the current state s_t. Taking the expectation of a_t conditional on h_{t-k} yields

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

where we used that $u_t^\varepsilon \perp\!\!\!\perp u_{t-k}^\varepsilon$ and $\mathbb{E}[u_t^\varepsilon] = 0$ by Assumption 3.2. As a result, the problem of learning π_h reduces to solving for π_h that satisfies the following identity

$$\mathbb{E}[a_t - \pi_h(h_t) \mid h_{t-k}] = 0, \tag{3}$$

which is a CMR problem as defined in Section 2. In this case, both a_t and h_t are observed in the confounded expert demonstrations, and h_{t-k} acts as the instrument.

To ensure that the instrument h_{t-k} is valid, we verify the three conditions from Section 2: firstly, $u_t^\varepsilon \perp\!\!\!\perp h_{t-k}$ as explained above. Secondly, the environment and expert policy are non-trivial since $\mathbb{P}(h_t \mid h_{t-k})$ is not constant in h_{t-k}, and, finally, h_{t-k} affects a_t only through s_t by the Markov property. However, the strength of h_{t-k}, representing its correlation with h_t, influences how well $\pi_h(h_t)$ can be identified. As the confounding horizon k increases, this correlation weakens, making h_{t-k} a less effective instrument. This relationship is formally analysed in Proposition 4.3 and validated through experiments in Section 5.

4.1 PRACTICAL ALGORITHMS FOR SOLVING THE CMRS

There are various techniques (Bennett et al., 2019a, Xu et al., 2020, Shao et al., 2024) for solving the CMRs $\mathbb{E}[a_t|h_{t-k}] = \mathbb{E}[\pi_h(h_t)|h_{t-k}]$. Here, the *CMR error* that we aim to minimise is given by

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

In Algorithm 1, we introduce DML-IL, an algorithm adapted from the IV regression algorithm DML-IV (Shao et al., 2024), which solves our CMRs by minimising the CMR error.[2] The first part of the algorithm (lines 3-7) learns a roll-out model $\hat{M}$ that generates a trajectory k steps ahead given

[2]DML stands for double machine learning (Chernozhukov et al., 2018), which is a statistical technique to ensure fast convergence rate for two-step regression, as is the case in Algorithm 1.

Algorithm 1 DML-IL

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

h_{t-k}. Then, $\hat{\pi}_h$ takes the generated trajectory $\hat{h}_t$ from $\hat{M}(h_{t-k})$ as input and minimises the mean square error to the next action (lines 8-13). Using generated trajectories is crucial in breaking the spurious correlation caused by u_t^ε, and the trajectory history before h_{t-k} allows the imitator to infer information about u_t^o. We refer to Appendix G for a discussion of the theoretical convergence rate guarantees of DML-IL and the choice of the confounding noise horizon k as input.

4.2 Theoretical Analysis

In this section, we derive theoretical guarantees for our algorithm, focusing on the imitation gap and its relationship to existing work. All proofs in this section are deferred to Appendix C.

In order to bound the imitation gap of the learnt policy $\hat{\pi}_h$, i.e., $J(\pi_E) - J(\hat{\pi}_h)$, we need to analyse:

(i) The information about hidden confounders u_t^o that can be inferred from trajectory histories;
(ii) The ill-posedness of the CMRs, which intuitively measures the strength of the instrument h_{t-k};
(iii) The disturbance of the confounding noise to the states and actions at test time.

These factors are all determined by the environment and the expert policy. To control (i), we measure how much information about u_t^o is captured by the trajectory history h_t by analysing the Total Variation (TV) distance between the distribution of u_t^o and $\mathbb{E}[u_t^o|h_t]$ along the trajectories of π_E. To control (ii) and (iii), we need to introduce the following two key concepts.

Definition 4.2 (The ill-posedness of CMRs (Dikkala et al., 2020))**.** Given the derived CMRs in Equation (3), the *ill-posedness* $\nu(\Pi, k)$ of the policy space with confounding noise horizon k is

$$\nu(\Pi, k) = \sup_{\pi \in \Pi} \frac{\|\pi_E - \pi\|_2}{\|\mathbb{E}[a_t - \pi(h_t)|h_{t-k}]\|_2}.$$

The ill-posedness $\nu(\Pi, k)$ measures the strength of the instrument, where a higher $\nu(\Pi, k)$ indicates a weaker instrument. As discussed previously, intuitively, the strength of the instrument would decrease as the confounding horizon k increases. This is confirmed by the following proposition.

Proposition 4.3. *$\nu(\Pi, k)$ is monotonically increasing as the confounded horizon k increases.*

Next, we introduce the notion of c-TV stability.

Definition 4.4 (c-total variation stability (Bassily et al., 2021, Swamy et al., 2022b))**.** Let $P(X)$ be the distribution of a random variable $X : \Omega \to \mathcal{X}$. $P(X)$ is c-TV stable if for $a_1, a_2 \in \mathcal{X}$ and $\Delta > 0$,

$$\|a_1 - a_2\| \leq \Delta \implies \delta_{TV}(a_1 + X, a_2 + X) \leq c\Delta$$

where $\|\cdot\|$ is some norm defined on $\mathcal{X}$ and δ_{TV} is the total variation distance.

A wide range of distributions are c-TV stable. For example, standard normal distributions are $\frac{1}{2}$-TV stable. We apply this notion to the distribution over u_t^ε to bound the disturbance it induces in the trajectory and the expected return.

With the notion of ill-posedness and c-TV stability, we can now analyse and upper bound the imitation gap $J(\pi_E) - J(\hat{\pi}_h)$ by controlling the three components $(i) - (iii)$ discussed above.

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*

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

This upper bound scales at the rate of T^2, which aligns with the expected behaviour of imitation learning without an interactive expert (Ross & Bagnell, 2010). Next, we show that the upper bounds of the imitation gap from prior work (Swamy et al., 2022b;a) 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 $J(\pi_E) - J(\hat{\pi}_h) \leq T^2\big(c\varepsilon\nu(\Pi, k)\big)$, which coincides with the bound of Theorem 5.1 in Swamy et al. (2022b).*

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 $J(\pi_E) - J(\hat{\pi}_h) \leq T^2\big(2\varepsilon/\sqrt{\dim(A)} + 2\delta\big)$, which is a concrete bound that extends the abstract bound in Theorem 5.4 of Swamy et al. (2022a).*

Remark 4.8. *If both u_t^ε and u_t^o are zero, we recover the classic setting of IL without confounders (Ross & Bagnell, 2010), and the imitation gap bound is $T^2\varepsilon$, where ε is the optimisation error of the algorithm.*

5 Experiments

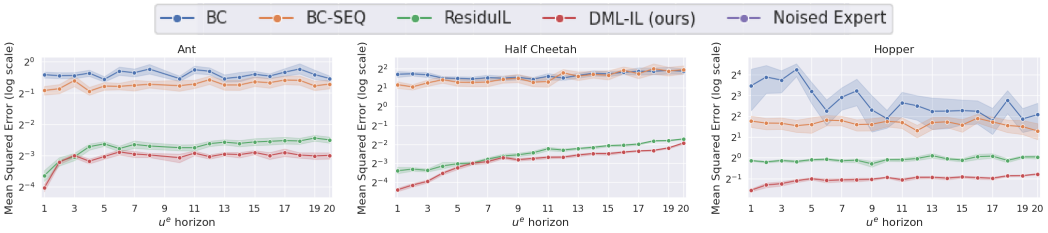
We here empirically evaluate the performance of Algorithm 1 (DML-IL) on Mujoco environments: Ant, Half Cheetah and Hopper. In our experiments, we compare with: Behavioural Cloning (BC), which learns $\mathbb{E}[a_t|s_t]$; BC-SEQ (Swamy et al., 2022a), which learns a history-dependent policy $\mathbb{E}[a_t|h_t]$; ResiduIL (Swamy et al., 2022b), which we adapt to our setting with h_{t-k} as instruments to learn a history-independent policy; and the noised expert, which is the maximally achievable performance.

We train imitators with 20000 samples (40 trajectories of 500 steps each) of the expert trajectory. The average reward is scaled such that 1 is the expert and 0 is a random policy. We also report the Mean Squared Error (MSE) between the imitator's and expert's actions. When the confounding noise u^ε is not specifically accounted for, we should expect to observe a much higher MSE. We vary the confounding noise horizon k from 1 to 20 in order to observe its effect on the strength of the instruments h_{t-k}. All results are plotted with one standard deviation as a shaded area. We also provide additional experiments in different environments in Appendix D. Moreover, we evaluate the use of other IV regression algorithms as the core CMR solver in Appendix D.2.

5.1 Mujoco Environments

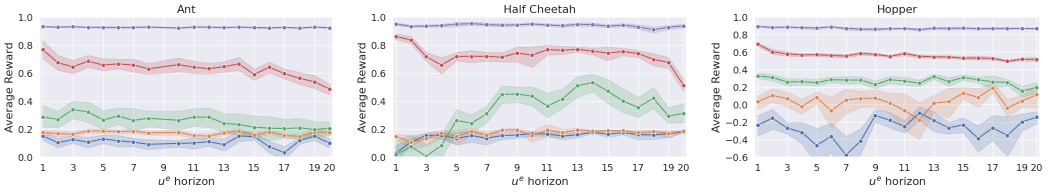
Experimental Setup. The original Mujoco environments do not have hidden variables, so we modify the environments and introduce expert-observable and expert-unobservable confounders. Specifically, while the original goal in Mujoco environments is to move forward as quickly as possible, we set the goal of travelling at a target speed u_t^o that varies throughout an episode. This varying target speed is observed by the expert but is not recorded in the dataset and acts as the expert-observable confounder u_t^o. In addition, we introduce additive confounding noise u_t^ε to states s_t and actions a_t to mimic confounding noise such as wind. Additional details about the modifications made to the environments are provided in Appendix E.2.

Results. We find that DML-IL consistently outperforms all other methods in terms of both MSE (Figure 2) and average reward (Figure 3), especially when the confounding horizon is 1. This suggests that DML-IL is successful in handling both types of confounders u_t^ε and u_t^o. ResiduIL is able to reduce the confounding effect of u_t^ε, which is evident by the comparatively low MSE (Figure 2). However, ResiduIL has no information about u_t^o and the best it can do is to assume an average (or expectation) of u_t^o, which nevertheless results in a worse average reward (Figure 3). Both BC and



(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.

BC-SEQ fail completely in the presence of confounding noise u_t^ε, with orders of magnitude higher MSE and average reward close to a random policy. From the similar performance of BC-SEQ and BC, we see that the use of trajectory histories to infer u_t^o is not helpful when the confounding noise u_t^ε is not handled explicitly. We also find that, as the confounding noise horizon k increases (x-axis), the MSE of DML-IL increases (Figure 2) as well as its average reward decreases (Figure 3). This corroborates the observation that the instrument is weaker and less information about u_t^o can be inferred from h_{t-k} as the confounding horizon k increases (see Proposition 4.3).

6 Conclusion

In this paper, we proposed a unifying framework for confounded IL with hidden confounders that unifies and extends previous confounded IL settings. Specifically, we considered hidden confounders to be partially observable to the expert, and demonstrated that causal IL under this framework can be reduced to a set of CMRs with the trajectory histories as instruments. We proposed DML-IL, a novel algorithm to solve these CMRs and learn an imitator. We provided bounds on the imitation gap for the learnt imitator. Finally, we empirically evaluated DML-IL on multiple tasks, including Mujoco environments, and demonstrated state-of-the-art performance against other causal IL algorithms.

Acknowledgments

This work was supported by the EPSRC Prosperity Partnership FAIR (grant number EP/V056883/1). DS acknowledges funding from the Turing Institute and Accenture collaboration. MK receives funding from the ERC under the European Union's Horizon 2020 research and innovation programme (FUN2MODEL, grant agreement No. 834115).

References

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455, 6 1996. ISSN 1537274X. doi: 10.1080/01621459.1996.10476902.

Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. *Uncertainty Proceedings 1994*, pp. 46–54, 1 1994. doi: 10.1016/B978-1-55860-332-5.50011-0.

Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *Robotics: Science and Systems*, 5:5986, 12 2018. ISSN 2330765X. doi: 10.15607/RSS.2019.XV.031. URL https://arxiv.org/abs/1812.03079v1.

Raef Bassily, Thomas Steinke, Kobbi Nissim, Uri Stemmer, Adam Smith, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. *SIAM Journal on Computing*, pp. 1046–1059, 11 2021. ISSN 07378017. doi: 10.1145/2897518.2897566. URL https://arxiv.org/abs/1511.02513v1.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in Neural Information Processing Systems*, 32, 2019a. ISSN 10495258.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in Neural Information Processing Systems*, 32, 2019b. ISSN 10495258.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. *Proceedings of Machine Learning Research*, 100:66–75, 12 2019. ISSN 26403498. doi: 10.1126/scirobotics.abc5986. URL https://arxiv.org/abs/1912.12294v1.

Xiaohong Chen and Timothy M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9:39–84, 3 2018. ISSN 17597331. doi: 10.3982/qe722.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. ISSN 1368-4221. doi: 10.1111/ECTJ.12097.

Sanjiban Choudhury, Mohak Bhardwaj, Sankalp Arora, Ashish Kapoor, Gireeja Ranade, Sebastian Scherer, and Debadeepta Dey. Data-driven planning via imitation learning. *International Journal of Robotics Research*, 37:1632–1672, 11 2017. ISSN 17413176. doi: 10.1177/0278364918781001. URL https://arxiv.org/abs/1711.06391v1.

Felipe Codevilla, Eder Santana, Antonio Lopez, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October:9328–9337, 4 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00942. URL https://arxiv.org/abs/1904.08980v1.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32, 5 2019. ISSN 10495258. URL https://arxiv.org/abs/1905.11979v2.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 2020-December, 6 2020. ISSN 10495258. URL https://arxiv.org/abs/2006.07201v1.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. *Proceedings of the 34th International Conference on Machine Learning*, 2017. doi: 10.5555/3305381.3305527.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in Neural Information Processing Systems*, pp. 4572–4580, 6 2016. ISSN 10495258. URL https://arxiv.org/abs/1606.03476v1.

Fateme Jamshidi, Sina Akbari, and Negar Kiyavash. Causal imitability under context-specific independence relations. 6 2023. URL https://arxiv.org/abs/2306.00585v2.

S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. *International Conference on Machine Learning*, 2002.

Alex Kuefler, Jeremy Morton, Tim Wheeler, and Mykel Kochenderfer. Imitating driver behavior with generative adversarial networks. *IEEE Intelligent Vehicles Symposium, Proceedings*, 5:204–211, 1 2017. doi: 10.1109/IVS.2017.7995721. URL https://arxiv.org/abs/1701.06699v1.

Daniel Kumor, Junzhe Zhang, and Elias Bareinboim. Sequential causal imitation learning with unobserved confounders. *Proceedings of the 35th Conference on Neural Information Processing Systems*, 8 2021. URL https://arxiv.org/abs/2208.06276v1.

Yann Lecun, Urs Muller, Jan Ben, Eric Cosatto, and Beat Flepp. Off-road obstacle avoidance through end-to-end learning. *Advances in Neural Information Processing Systems*, 18, 2005. URL http://yann.lecun.com.

Luofeng Liao, You Lin Chen, Zhuoran Yang, Bo Dai, Zhaoran Wang, and Mladen Kolar. Provably efficient neural estimation of structural equation model: An adversarial approach. *Advances in Neural Information Processing Systems*, 2020-December, 7 2020. ISSN 10495258. URL https://arxiv.org/abs/2007.01290v3.

Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71:1565–1578, 9 2003. ISSN 1468-0262. doi: 10.1111/1468-0262.00459.

Pedro A. Ortega and Daniel A. Braun. A minimum relative entropy principle for learning and acting. *Journal of Artificial Intelligence Research*, 38:475–511, 10 2008. ISSN 10769757. doi: 10.1613/jair.3062. URL https://arxiv.org/abs/0810.3605v3.

Pedro A. Ortega, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, Tom Everitt, Corentin Tallec, Emilio Parisotto, Tom Erez, Yutian Chen, Scott Reed, Marcus Hutter, Nando de Freitas, and Shane Legg. Shaking the foundations: delusions in sequence models for interaction and control. 10 2021. URL https://arxiv.org/abs/2110.10819v1.

Judea Pearl. Causality: Models, reasoning, and inference. *Econometric Theory*, 2000.

Samuel Pfrommer, Yatong Bai, Hyunin Lee, and Somayeh Sojoudi. Initial state interventions for deconfounded imitation learning. 7 2023. URL https://arxiv.org/abs/2307.15980v3.

Dean A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in Neural Information Processing Systems*, 1, 1988.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Stéphane Ross and J Andrew Bagnell. Efficient reductions for imitation learning. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. *Journal of Machine Learning Research*, 15: 627–635, 11 2011. ISSN 15324435.

Kangrui Ruan, Junzhe Zhang, Xuan Di, and Elias Bareinboim. Causal imitation learning via inverse reinforcement learning. *Proceedings at the International Conference on Learning Representations*, 2023.

Stuart Russell. Learning agents for uncertain environments (extended abstract). *In The Eleventh Annual Conference on Computational Learning Theory*, 1998.

Daqian Shao, Ashkan Soleymani, Francesco Quinzan, and Marta Kwiatkowska. Learning decision policies with instrumental variables through double machine learning. *Proceedings of the International Conference on Machine Learning*, 2024.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008. ISSN 1533-7928. URL http://jmlr.org/papers/v9/shpitser08a.html.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 6 2019. ISSN 10495258.

Jonathan Spencer, Sanjiban Choudhury, Arun Venkatraman, Brian Ziebart, and J. Andrew Bagnell. Feedback in imitation learning: The three regimes of covariate shift. 2 2021. URL https://arxiv.org/abs/2102.02872v2.

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, and Zhiwei Steven Wu. Sequence model imitation learning with unobserved contexts. *Advances in Neural Information Processing Systems*, 35, 8 2022a. ISSN 10495258.

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, and Zhiwei Steven Wu. Causal imitation learning under temporally correlated noise. *Proceedings of Machine Learning Research*, 162:20877–20890, 2 2022b. ISSN 26403498. URL https://arxiv.org/abs/2202.01312v1.

Risto Vuorio, Johann Brehmer, Hanno Ackermann, Daniel Dijkman, Taco Cohen, and Pim de Haan. Deconfounded imitation learning. 11 2022. URL https://arxiv.org/abs/2211.02667v1.

Chuan Wen, Jierui Lin, Trevor Darrell, Dinesh Jayaraman, and Yang Gao. Fighting copycat agents in behavioral cloning from observation histories. *Advances in Neural Information Processing Systems*, 2020-December, 10 2020. ISSN 10495258. URL https://arxiv.org/abs/2010.14876v1.

Philip G. Wright. The tariff on animal and vegetable oils. *https://doi.org/10.1086/254144*, 38: 619–620, 10 1928. ISSN 0022-3808. doi: 10.1086/254144.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. *ICLR 2021 - 9th International Conference on Learning Representations*, 10 2020. URL https://arxiv.org/abs/2010.07154v4.

Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. *Proceedings of the 34th Conference on Neural Information Processing Systems*, 8 2020. URL https://arxiv.org/abs/2208.06267v1.

A Related Work

Imitation Learning. Imitation learning considers the problem of learning from demonstrations (Pomerleau, 1988, Lecun et al., 2005). Standard IL methods include Behaviour Cloning (Pomerleau, 1988), Inverse RL (Russell, 1998), and adversarial methods (Ho & Ermon, 2016). Interactive IL (Ross et al., 2011) extends standard IL by allowing the imitator to query an interactive expert, facilitating recovery from mistakes. However, in this paper, we do not assume query access to an interactive expert.

Causal Imitation Learning. Recently, it has been shown that IL from offline trajectories can suffer from the existence of latent variables (Ortega et al., 2021), which cause causal delusion. This can be resolved by learning an interventional policy. Following this discovery, various methods (Vuorio et al., 2022, Swamy et al., 2022a) considered IL when the expert has access to the full hidden context that is fixed throughout each episode, but the imitator does not observe the hidden context. They aim to learn an interventional policy through on-policy IL algorithms that require an interactive demonstrator and/or an interactive simulator (e.g., DAgger (Ross et al., 2011)).

Orthogonal to these works, Swamy et al. (2022b) consider latent variables not known to the expert, which act as confounding noise that affects the expert policy, but not the transition dynamics. To address this challenge, the problem is then cast into an IV regression problem. Our work combines and generalises the above works (Vuorio et al., 2022, Swamy et al., 2022a;b) to allow the latent variables to be only partly known to the expert, evolving through time in each episode and directly affecting both the expert policy and the transition dynamics. Solving this generalisation implies solving the above problems simultaneously.

Causal confusion (de Haan et al., 2019, Pfrommer et al., 2023) considers the situation where the expert's actions are spuriously correlated with non-causal features of the previous observable states. While it is implicitly assumed that there are no latent variables present in the environment, we can still model this spurious correlation as the existence of hidden confounders that affect both previous states and current expert actions. Slight variations of this setting have been studied in (Wen et al., 2020, Spencer et al., 2021, Codevilla et al., 2019). In Appendix B, we explain and discuss how these works can be reduced to special cases of our unifying framework.

From the causal inference perspective (Kumor et al., 2021, Zhang et al., 2020), there have been studies of the theoretical conditions on the causal graph such that the imitator can exactly match the expert performance through backdoor adjustments (*imitability*). Similarly, Ruan et al. (2023) extended imitation conditions and backdoor adjustments to inverse RL. We instead consider a setting where exact imitation is not possible and aim to minimise the imitation gap. Beyond backdoor adjustments, imitability has also been studied theoretically using context-specific independence relations (Jamshidi et al., 2023).

IV Regression and CMRs. In this paper, we transform our causal IL problem into solving a set of CMRs through IVs. Therefore, we briefly introduce IV regression and approaches for solving CMRs. The classic IV regression algorithms mainly consider linear functions (Angrist et al., 1996) and non-linear basis functions (Newey & Powell, 2003, Chen & Christensen, 2018, Singh et al., 2019). More recently, DNNs have been used for function estimation and methods such as DeepIV (Hartford et al., 2017), DeepGMM (Bennett et al., 2019b), AGMM (Dikkala et al., 2020), DFIV (Xu et al., 2020) and DML-IV (Shao et al., 2024) have been proposed.

More generally, IV regression algorithms can be generalised to solve CMRs (Liao et al., 2020, Dikkala et al., 2020, Shao et al., 2024), specifically linear CMRs, where the restrictions are linear functionals of the function of interest. In our paper, the derived CMRs for causal IL are linear, so the above methods can be adopted.

B Reducing our unifying framework to related literature

In this section, we discuss how the various previous works can be obtained as special cases of our unifying framework.

B.1 TEMPORALLY CORRELATED NOISE (SWAMY ET AL., 2022B)

The Temporally Correlated Noise (TCN) proposed in (Swamy et al., 2022b) is a special case of our setting where $u^o = 0$ and only the confounding noise u^ε is present. Following Equation 14-17 of (Swamy et al., 2022b), their setting can be summarised as

$$\begin{aligned} s_t &= \mathcal{T}(s_{t-1}, a_{t-1}) \\ &= \mathcal{T}(s_{t-1}, \pi_E(s_{t-1}) + u_{t-1} + u_{t-2}) \\ a_t &= \pi_E(s_t) + u_t + u_{t-1}, \end{aligned}$$

where $\mathcal{T}$ is the transition function and u_t are the TCN. It can be seen that TCN is the confounding noise u^ε since the expert policy doesn't take it into account and it affects (or confounds) both the state and action.

It can be seen that this is a special case of our framework when $u_t^o = 0$, where $a_t = \pi_E(s_t) + \varepsilon(u_t^\varepsilon)$ from Equation (2), and more specifically when the confounding noise horizon in Theorem 3.2 is 2. In addition, the theoretical results in (Swamy et al., 2022b) can be deduced from our main results as shown in Corollary 4.7.

B.2 UNOBSERVED CONTEXTS (SWAMY ET AL., 2022A)

The setting considered by Swamy et al. (2022a) is a special case of our setting when $u^\varepsilon = 0$ and only u^o are present. Following Section 3 of Swamy et al. (2022a), their setting can be summarised as

$$\begin{aligned} &\mathcal{T} : \mathcal{S} \times \mathcal{A} \times C \to D(S) \\ &\nabla : \mathcal{S} \times \mathcal{A} \times C \to [-1, 1] \\ &a_t = \pi_E(s_t, c) \end{aligned}$$

where $c \in C$ is the context, which is assumed to be fixed throughout an episode. There are no hidden confounders in this setting and the context c is included in u^o under our framework. Note that in our setting we also allow u^o to be varying throughout an episode. In addition, the theoretical results in (Swamy et al., 2022a) can be deduced from our main results as shown in Corollary 4.6.

B.3 IMITATION LEARNING WITH LATENT CONFOUNDERS (VUORIO ET AL., 2022)

The setting considered by (Vuorio et al., 2022) is also a special case of our setting when $u^\varepsilon = 0$ and only u^o are present, which is very similar to (Swamy et al., 2022a). In Section 2.2 of (Vuorio et al., 2022), they introduced a latent variable $\theta \in \Theta$ that is fixed throughout an episode and $a_t = \pi_E(s_t, \theta)$. There are no hidden confounders in this setting and the latent variable θ is included in u^o in our framework. No theoretical imitation gap bounds are provided in Vuorio et al. (2022). However, Corollary 4.6 can be directly applied to their setting and bound the imitation gap.

B.4 CAUSAL DELUSION AND CONFUSION DE HAAN ET AL. (2019), WEN ET AL. (2020), ORTEGA ET AL. (2021), SPENCER ET AL. (2021), PFROMMER ET AL. (2023)

The concept of causal delusion (Ortega et al., 2021) and confusion is widely studied in the literature (de Haan et al., 2019, Wen et al., 2020, Spencer et al., 2021, Pfrommer et al., 2023) from different perspectives. A classic example of causal confusion is learning to break in an autonomous driving scenario. The states are images with full view of the dashboard and the road conditions. The break indicator in this scenario is the confounding variable that correlates with the action of breaking in subsequent steps, which causes the imitator to learn to break if the break indicator light is already on. Therefore, another name for this problem is the latching problem, where the imitator latches to spurious correlations between current action and the trajectory history.

In the setting of Ortega et al. (2021), this is explicitly modelled as latent variables that affect both the action and state, causing spurious correlation between them and confusing the imitator. In other settings de Haan et al. (2019), Pfrommer et al. (2023), Spencer et al. (2021), Wen et al. (2020), there are no explicit unobserved confounders, but the nuisance correlation between the previous states and actions can be modelled as the existence of hidden confounders u^ε in our framework. Specifically, in de Haan et al. (2019), x_{t-1} and a_{t-1} are considered confounders that affect the state variable

x_t, which causes a spurious correlation between previous state action pairs and a_t. The spurious correlation between variables is typically modelled as the existence of a hidden confounder u^ε that affects both variables in causal modelling. For example, the actual hazard or event that causes the expert to break will be the hidden confounder u^ε that affects both the break and the break indicator.

However, despite the fact that this setting can be considered a special case of our general framework, we stress that the concrete and practical problems considered in de Haan et al. (2019), Pfrommer et al. (2023), Spencer et al. (2021), Wen et al. (2020) are different from ours, where they assumed implicitly that the hidden confounders u^ε are embedded in the observations or outright observed.

C Proofs of Main Results

In this section, we provide the proofs for the main results and corollaries in this paper.

C.1 Proof of Propositions

Proposition 4.3: The ill-posedness $\nu(\Pi, k)$ is monotonically increasing as the confounded horizon k increases.

Proof. From definition, we have that

$$\nu(\Pi, k) = \sup_{\pi \in \Pi} \frac{\|\pi_E - \pi\|_2}{\|\mathbb{E}[a_t - \pi(h_t)|h_{t-k}]\|_2}.$$

We would like to show for each $\pi \in \Pi$, $\frac{\|\pi_E - \pi\|_2}{\|\mathbb{E}[a_t - \pi(h_t)|h_{t-k}]\|_2}$ is increasing as k increases, which would imply that $\nu(\Pi, k)$ is increasing. For each $\pi \in \Pi$, we see that the numerator is constant w.r.t the horizon k. Therefore, it is enough to check that for each $\pi \in \Pi$, the denominator $\|\mathbb{E}[a_t - \pi(h_t)|h_{t-k}]\|_2$ decreases as k increases. For any two integer horizon $k_1 > k_2$,

$$\begin{aligned}
\mathbb{E}[a_t - \pi(h_t)|h_{t-k_1}]^2 &= \mathbb{E}[\mathbb{E}[a_t - \pi(h_t)|h_{t-k_2}]|h_{t-k_1}]^2 &(4)\\
&\leq \mathbb{E}[\mathbb{E}[a_t - \pi(h_t)|h_{t-k_2}]^2|h_{t-k_1}] &(5)\\
&= \mathbb{E}[a_t - \pi(h_t)|h_{t-k_2}]^2 &(6)
\end{aligned}$$

by the tower property of conditional expectation as $\sigma(h_{t-k_1}) \subseteq \sigma(h_{t-k_2})$, Jensen's inequality for conditional expectations, and the fact that $\mathbb{E}[a_t - \pi(h_t)|h_{t-k_2}]^2$ is h_{t-k_1} measurable, respectively for each line. Therefore, we have that $\mathbb{E}[a_t - \pi(h_t)|h_{t-k}]$ is decreasing, which implies $\|\mathbb{E}[a_t - \pi(h_t)|h_{t-k}]\|_2$ is decreasing and $\nu(\Pi, k)$ is increasing as k increases, which completes the proof. □

C.2 Main results for guarantees on the imitation gap

Theorem 4.5: 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

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

Proof of Theorem 4.5. We let the Q-function of a policy $\pi_h \in \Pi$ be defined as $Q_\pi(s_t, a_t, u_t^o) = \mathbb{E}_{\tau \sim \pi_h}[\sum_{t'=t}^{T} r(s_{t'}, a_{t'}, u_{t'}^o)]$. Recall that $J(\pi)$ is the expected reward following π, and we would like to bound the performance gap $J(\pi_E) - J(\hat{\pi_h})$ between the expert policy π_E and the learned history-dependent policy $\hat{\pi_h}$. Let $Q_{\hat{\pi_h}}(s_t, a_t, u_t^o)$ be the Q-function of $\hat{\pi}_h$. Using the Performance Difference Lemma (Kakade & Langford, 2002), we have that for any Q-function $\tilde{Q}(h_t, a_t)$ that takes

in the trajectory history h_t and action a_t,

$$\begin{aligned}
J(\pi_E) - J(\hat{\pi}_h) &= \mathbb{E}_{\tau\sim\pi_E}[\sum_{t=1}^{T} Q_{\hat{\pi}_h}(s_t, a_t, u_t^o) - \mathbb{E}_{a\sim\hat{\pi_h}}[Q_{\hat{\pi_h}}(s_t, a, u_t^o)]] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_E}[Q_{\hat{\pi_h}}(s_t, a_t, u_t^o) - \tilde{Q}(h_t, a_t) + \tilde{Q}(h_t, a_t) - \mathbb{E}_{a\sim\hat{\pi_h}}[Q_{\hat{\pi_h}} - \tilde{Q} + \tilde{Q}]] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_E}[\tilde{Q} - \mathbb{E}_{a\sim\hat{\pi_h}}[\tilde{Q}]] + \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_E}[Q_{\hat{\pi_h}} - \tilde{Q} - \mathbb{E}_{a\sim\hat{\pi_h}}[Q_{\hat{\pi_h}} - \tilde{Q}]] \quad (7)
\end{aligned}$$

We first bound the second part of Equation (7). Denote by δ_{TV} the total variation distance. For two distributions P, Q, recall the property of total variation distance for bounding the difference in expectations:

$$|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]| \leq \|f\|_\infty \delta_{TV}(P, Q).$$

In order to bound the second part of Equation (7), for any Q function, consider inferred $\tilde{Q}$ using the conditional expectation of u^o on the history h,

$$\tilde{Q}(h_t, a_t) \coloneqq Q(s_t, a_t, \mathbb{E}_{\tau\sim\pi_E}[u_t^o|h_t]),$$

where note that $s_t \in h_t$. We have that, when the transition trajectory $(s_t, u_t^o, u_t^\varepsilon, r_t) \sim \pi_E$ follows the expert policy, for any action $\dot{a} \sim \pi$ following some policy π (in our case, it can be π_E or $\hat{\pi_h}$),

$$\begin{aligned}
&|\mathbb{E}_{\tau\sim\pi_E, \dot{a}\sim\pi}[Q(s_t, \dot{a}, u_t) - \tilde{Q}(h_t, \dot{a})]| && (8)\\
&= |\mathbb{E}_{\tau\sim\pi_E, \dot{a}\sim\pi}[Q(s_t, \dot{a}, u_t^o) - Q(s_t, \dot{a}, \mathbb{E}_{\tau\sim\pi_E}[u_t^o|h_t])]| \\
&= \left|\mathbb{E}_{u_t^o\sim\pi_E}[\mathbb{E}_{\pi_E,\pi}[Q(s_t, \dot{a}, u_t^o)|u_t^o] - \mathbb{E}_{u_t^o|h_t\sim\pi_E}[\mathbb{E}_{\pi_E,\pi}[Q(s_t, \dot{a}, u_t^o)|u_t^o]\right| && (9)\\
&\leq \|\mathbb{E}_{\pi_E,\pi}[Q(s_t, \dot{a}, u_t^o)|u_t^o]\|_\infty \delta_{TV}(u_t^o, \mathbb{E}_{\pi_E}[u_t^o|h_t]) && (10)\\
&\leq T \cdot \delta_{TV}(u_t^o, \mathbb{E}_{\pi_E}[u_t^o|h_t]) && (11)\\
&\leq T\delta && (12)
\end{aligned}$$

where Equation (9) uses the tower property of expectations, Equation (10) uses the total variation distance bound for bounded functions, Equation (11) uses the fact that the Q function is bounded by T and Equation (12) uses the condition that $\delta_{TV}(u_t^o, \mathbb{E}_{\pi_E}[u_t^o|h_t]) \leq \delta$ in the theorem statement. Since Equation (7) holds for any choice of $\tilde{Q}$, we choose $\tilde{Q}_{\hat{\pi_h}}(h_t, a_t) \coloneqq Q_{\hat{\pi_h}}(s_t, a_t, \mathbb{E}_{\tau\sim\pi_E}[u_t^o|h_t])$ such that we can apply Equation (12) twice to bound the second part of Equation (7):

$$\begin{aligned}
\mathbb{E}_{\tau\sim\pi_E}[Q_{\hat{\pi_h}} - \tilde{Q}_{\hat{\pi_h}} - \mathbb{E}_{a\sim\hat{\pi_h}}[Q_{\hat{\pi_h}} - \tilde{Q}_{\hat{\pi_h}}]] && (13)\\
&\leq \mathbb{E}_{\tau\sim\pi_E}[Q_{\hat{\pi_h}} - \tilde{Q}_{\hat{\pi_h}} + |\mathbb{E}_{a\sim\hat{\pi_h}}[Q_{\hat{\pi_h}} - \tilde{Q}_{\hat{\pi_h}}]|] \\
&= \mathbb{E}_{\tau\sim\pi_E}[Q_{\hat{\pi_h}} - \tilde{Q}_{\hat{\pi_h}}] + |\mathbb{E}_{s_t,u_t\sim\pi_E, a\sim\hat{\pi_h}}[Q_{\hat{\pi_h}} - \tilde{Q}_{\hat{\pi_h}}]| \\
&\leq |\mathbb{E}_{\tau\sim\pi_E}[Q_{\hat{\pi_h}} - \tilde{Q}_{\hat{\pi_h}}]| + T\delta && (14)\\
&\leq 2T\delta
\end{aligned}$$

where Equation (14) holds by applying Equation (12) because the expectation of the trajectories (and their transitions) are over π_E, and the actions which are used only as arguments into the Q function are sampled from $\hat{\pi_h}$.

Next, we bound the first part of Equation (7). Recall that the ill-posedness of the problem for a policy class Π is

$$\nu(\Pi, k) = \sup_{\pi\in\Pi} \frac{\|\pi_E - \pi\|_2}{\|\mathbb{E}[a_t - \pi(h_t)|h_{t-k}]\|_2}$$

where $\|\pi_E - \pi\|_2$ is the RMSE and $\|\mathbb{E}[a_t - \pi(s_t)|s_{t-k}]\|_2$ is the CMR error from our algorithm. Since the learned policy $\hat{\pi_h}$ have CMR error of ε, we have that

$$\|\pi_E - \hat{\pi_h}\|_2 \leq \nu(\Pi, k)\|\mathbb{E}[a_t - \hat{\pi_h}(h_t)|h_{t-k}]\|_2 \leq \nu(\Pi, k)\varepsilon$$

Next, recall that c-total variation stability of a distribution $P(u^\varepsilon)$ where $u^\varepsilon \in A$ for some space A implies for two elements $a_1, a_2 \in A$,

$$\|a_1 - a_2\|_2 \leq \Delta \implies \delta_{TV}(a_1 + u^\varepsilon, a_2 + u^\varepsilon) \leq c\Delta.$$

Since $P(u_t^\varepsilon)$ is c-TV stable w.r.t the action space A, we have that for all history trajectories $h_t \in H$ (note that $s_t \in h_t$)

$$\delta_{TV}(\pi_E(s_t) + u_t^\varepsilon, \hat{\pi_h}(h_t) + u_t^\varepsilon) \leq c\|\pi_E(s_t) - \hat{\pi_h}(h_t)\|_2.$$

Then, we have that by Jensen's inequality,

$$\mathbb{E}_{h_t\sim\pi_E}[\delta_{TV}(\pi_E(s_t) + u_t^\varepsilon, \hat{\pi_h}(h_t) + u_t^\varepsilon)]^2 \leq \mathbb{E}_{h_t\sim\pi_E}[\delta_{TV}(\pi_E(s_t) + u_t^\varepsilon, \hat{\pi_h}(h_t) + u_t^\varepsilon)^2]$$

Consequently, we obtain

$$\begin{aligned}\mathbb{E}_{h_t\sim\pi_E}[\delta_{TV}(\pi_E(s_t) + u_t^\varepsilon, \hat{\pi_h}(h_t) + u_t^\varepsilon)] &\leq \sqrt{\mathbb{E}_{h_t\sim\pi_E}[\delta_{TV}(\pi_E(s_t) + u_t^\varepsilon, \hat{\pi_h}(h_t) + u_t^\varepsilon)^2]} \\ &\leq \sqrt{c^2\mathbb{E}_{h_t\sim\pi_E}[\|\pi_E(s_t) - \hat{\pi_h}(h_t)\|_2^2]} \\ &= c\|\pi_E - \hat{\pi_h}\|_2 \leq c\varepsilon\nu(\Pi, k)\end{aligned}$$

Therefore, by applying the total variation distance bound for expectations of $\tilde{Q}_{\hat{\pi_h}}$ over different distributions of action a_t, we have that

$$\mathbb{E}_{\tau\sim\pi_E}[\tilde{Q}_{\hat{\pi_h}} - \mathbb{E}_{a\sim\hat{\pi_h}}[\tilde{Q}_{\hat{\pi_h}}]] = \mathbb{E}_{\tau\sim\pi_E}[\tilde{Q}_{\hat{\pi_h}}(h_t, a_t) - \mathbb{E}[\tilde{Q}_{\hat{\pi_h}}(h_t, \hat{\pi_h}(h_t))]] \tag{15}$$

$$= \mathbb{E}_{h_t\sim\pi_E}[\mathbb{E}[\tilde{Q}_{\hat{\pi_h}}(h_t, \pi_E(s_t) + u_t^\varepsilon)] - \mathbb{E}[\tilde{Q}_{\hat{\pi_h}}(h_t, \hat{\pi_h}(h_t) + u_t^\varepsilon)]] \tag{16}$$

$$\leq \|\tilde{Q}_{\hat{\pi_h}}\|_\infty \mathbb{E}_{h_t\sim\pi_E}[\delta_{TV}(F(\pi_E(s_t) + u_t^\varepsilon), F(\hat{\pi_h}(h_t) + u_t^\varepsilon))] \tag{17}$$

$$\leq Tc\varepsilon\nu(\Pi, k) \tag{18}$$

Combining all of above, we see that from Equation (7), by selecting $\tilde{Q}_{\hat{\pi_h}}(h_t, a_t) := Q_{\hat{\pi_h}}(s_t, a_t, \mathbb{E}_{\tau\sim\pi_E}[u_t^o|h_t])$, the imitation gap can be bounded by

$$J(\pi_E) - J(\hat{\pi_h}) \tag{19}$$

$$= \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_E}[\tilde{Q}_{\hat{\pi_h}} - \mathbb{E}_{a\sim\hat{\pi_h}}[\tilde{Q}_{\hat{\pi_h}}]] + \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_E}[Q_{\hat{\pi_h}} - \tilde{Q}_{\hat{\pi_h}} - \mathbb{E}_{a\sim\hat{\pi_h}}[Q_{\hat{\pi_h}} - \tilde{Q}_{\hat{\pi_h}}]] \tag{20}$$

$$\leq \sum_{t=1}^{T} Tc\varepsilon\nu(\Pi, k) + \sum_{t=1}^{T} 2T\delta \tag{21}$$

$$\leq T \cdot (Tc\varepsilon\nu(\Pi, k) + 2T\delta) \tag{22}$$

$$= T^2(c\varepsilon\nu(\Pi, k) + 2\delta) = \mathcal{O}(T^2(\varepsilon + \delta)), \tag{23}$$

which concludes the proof. □

C.3 Proofs of Corollaries

Corollary 4.6: In the special case that $u_t^o = 0$, meaning that there is no confounder observable to the expert, or $u_t^o = \mathbb{E}_{\pi_E}[u_t^o|h_t]$, meaning that u_t^o is $\sigma(h_t)$ measurable (all information regarding u_t^o is represented in the history), the imitation gap bound is $T^2(c\varepsilon\nu(\Pi, k))$, which coincides with Theorem 5.1 of Swamy et al. (2022b).

Proof. If $u_t^o = 0$, then we have $u_t^o = \mathbb{E}_{\pi_E}[u_t^o|h_t]$ since u_t^o is a constant. If $u_t^o = \mathbb{E}_{\pi_E}[u_t^o|h_t]$, we have that

$$\delta_{TV}(u_t^o, \mathbb{E}_{\pi_E}[u_t^o|h_t]) = \delta_{TV}(u_t^o, u_t^o) \leq 0$$

By plugging $\delta = 0$ into Theorem 4.5, we have that $J(\pi_E) - J(\hat{\pi_h}) \leq T^2(c\varepsilon\nu(\Pi, k))$, which is the same as the imitation gap derived in Swamy et al. (2022b) and completes the proof. □

Corollary 4.7: In the special case that $u_t^\varepsilon = 0$, if the learnt policy via supervised BC have error ε, then the imitation gap bound is $T^2(\frac{2}{\sqrt{\dim(A)}}\varepsilon + 2\delta)$, which is a concrete bound that extends the abstract bound in Theorem 5.4 of Swamy et al. (2022a).

Proof. In Theorem 5.4 of Swamy et al. (2022a), for the offline case, which is the setting we are considering (as opposed to the online settings), they defined the following quantities for bounding the imitation gap in a very general fashion,

$$\varepsilon_{\text{off}} \coloneqq \sup_{\tilde{Q}} \mathbb{E}_{\tau\sim\pi_E}[\tilde{Q} - \mathbb{E}_{a\sim\hat{\pi_h}}[\tilde{Q}]]$$
$$\delta_{\text{off}} \coloneqq \sup_{Q\times\tilde{Q}} \mathbb{E}_{\tau\sim\pi_E}[Q_{\hat{\pi_h}} - \tilde{Q} - \mathbb{E}_{a\sim\hat{\pi_h}}[Q_{\hat{\pi_h}} - \tilde{Q}]].$$

The imitation gap by Theorem 5.4 in Swamy et al. (2022a) under the assumption that $u_t^\varepsilon = 0$ is $T^2(\varepsilon_{\text{off}} + \delta_{\text{off}})$, which can also be deduced from Equation (7) by naively applying the above supremum over all possible Q functions. To obtain a concrete bound, we can provide a tighter bound for $\mathbb{E}_{\tau\sim\pi_E}[\tilde{Q}_{\hat{\pi_h}} - \mathbb{E}_{a\sim\hat{\pi_h}}[\tilde{Q}_{\hat{\pi_h}}]]$, which is the first part of Equation (7), given that $u_t^\varepsilon = 0$.

For two elements $a_1, a_2 \in A$, we have that by Cauchy–Schwarz,

$$\delta_{TV}(a_1 + 0, a_2 + 0) = \frac{1}{2}\|a1 - a2\|_1 \leq \frac{\sqrt{\dim(A)}}{2}\|a1 - a2\|_2.$$

Then, we have that

$$\|a_1 - a_2\|_2 \leq \Delta \implies \delta_{TV}(a_1, a_2) \leq \frac{2}{\sqrt{\dim(A)}}\Delta$$

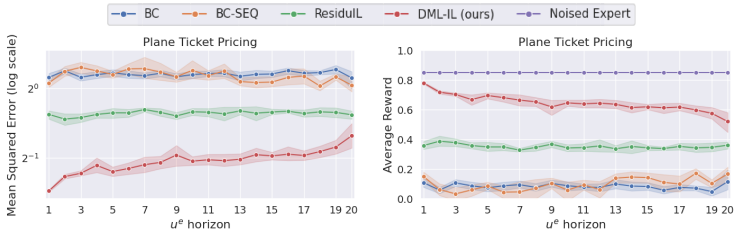
so that by Theorem 4.5,

$$\begin{aligned}
\mathbb{E}_{\tau\sim\pi_E}[\tilde{Q}_{\hat{\pi_h}} - \mathbb{E}_{a\sim\hat{\pi_h}}[\tilde{Q}_{\hat{\pi_h}}]] &= \mathbb{E}_{\tau\sim\pi_E}[\tilde{Q}_{\hat{\pi_h}}(h_t, a_t) - \mathbb{E}[\tilde{Q}_{\hat{\pi_h}}(h_t, \hat{\pi_h}(h_t))]] && (24)\\
&= \mathbb{E}_{h_t\sim\pi_E}[\mathbb{E}[\tilde{Q}_{\hat{\pi_h}}(h_t, \pi_E(s_t))] - \mathbb{E}[\tilde{Q}_{\hat{\pi_h}}(h_t, \hat{\pi_h}(h_t))]] && (25)\\
&\leq \|\tilde{Q}_{\hat{\pi_h}}\|_\infty \frac{2}{\sqrt{\dim(A)}}\|\pi_E - \pi\|_2 && (26)\\
&\leq T\frac{2}{\sqrt{\dim(A)}}\varepsilon, && (27)
\end{aligned}$$

since when $u_t^\varepsilon = 0$ the learning error via supervised learning is $\varepsilon := \|\pi_E - \pi\|_2$. Therefore, the final imitation bound following Theorem 4.5 is

$$\begin{aligned}
&J(\pi_E) - J(\hat{\pi_h}) && (28)\\
&= \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_E}[\tilde{Q}_{\hat{\pi_h}} - \mathbb{E}_{a\sim\hat{\pi_h}}[\tilde{Q}_{\hat{\pi_h}}]] + \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_E}[Q_{\hat{\pi_h}} - \tilde{Q}_{\hat{\pi_h}} - \mathbb{E}_{a\sim\hat{\pi_h}}[Q_{\hat{\pi_h}} - \tilde{Q}_{\hat{\pi_h}}]] && (29)\\
&\leq \sum_{t=1}^{T} T\frac{2}{\sqrt{\dim(A)}}\varepsilon + \sum_{t=1}^{T} 2T\delta && (30)\\
&= T^2(\frac{2}{\sqrt{\dim(A)}}\varepsilon + 2\delta). && (31)
\end{aligned}$$

This bound is a concrete bound, obtained through a detailed analysis of the problem at hand, that coincides with the abstract bound $T^2(\varepsilon_{\text{off}} + \delta_{\text{off}})$ provided in Theorem 5.4 of Swamy et al. (2022b). Note that this bound is independent of the ill-posedness $\nu(\Pi, k)$ and the c-TV stability of u_t^ε, which are present in the bound of Theorem 4.5, due to the lack of hidden confounders u_t^ε. □



(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 (Example 3.1).

D Additional Experiments

Here, we discuss a motivating example to illustrate the importance of considering $u_t = (u_t^o, u_t^\varepsilon)$.

Example D.1. Consider an airline ticket pricing scenario Wright (1928), where the goal is to learn a pricing policy by imitating actual airline pricing based on expert-set profit margins. Suppose that seasonal patterns and external events are known only to experts, but missing from the dataset, serving as expert-observable confounders u_t^o. Meanwhile, actual airline prices are confounded (additively) by fluctuating operating costs, which are unknown to the experts when they set the profit margin and unobserved in the dataset, making them confounding noise u_t^ε. We conduct experiments on a toy environment inspired by this in Appendix D.1 and show that IL algorithms that do not distinguish between u_t^o and u_t^ε fail to correctly imitate the expert.

D.1 Airline Ticket Pricing Environment

Experimental Setup. We conducted additional experiments on the airline ticket pricing environment described in Example 3.1. The confounding noise u^ε are operation costs and u_t^o are seasonal demand patterns and events. Details on this environment are provided in Appendix E.1.

Results. The results are presented in Figure 4. DML-IL performed the best with the lowest MSE and the highest average reward that is closest to the expert, especially when the u_t^ε horizon is 1. Overall, we observe very similar results to those from the Mujoco evironments in Section 5.

D.2 Adopting other IV regression algorithms

In this paper, we have transformed causal IL with hidden confounders into a set of CMRs as defined in Equation (3). Therefore, in principle many IV regression algorithms can be adopted to solve our CMRs. We also experimented with other IV regression algorithms that have been previously shown to be practical Shao et al. (2024) for different tasks and high-dimensional input. Specifically, we experimented with DFIV Xu et al. (2020), which is an iterative algorithm that integrates the training of two models that depend on each other, and DeepGMM Bennett et al. (2019b), which solves a minimax game by optimising two models adversarially. Note that DeepIV Hartford et al. (2017) can be considered a special case of DML-IV Shao et al. (2024), so we did not reimplement it.

The additional results for using DFIV and DeepGMM as the CMRs solver are provided in Figure 5 and Figure 6. It can be seen from Figure 5 that only DFIV achieves good performance in the airline ticket pricing environment, surpassing the performance of ResiduIL. For the Ant Mujoco task in Figure 6, both DFIV and DeepGMM fail to learn good policies, with only slightly lower MSE than BC and BC-SEQ. We think this is mainly due to the high-dimensional state and action spaces and the inherent instability in the DFIV and DeepGMM algorithms. For DFIV, the interleaving of training of two models causes highly non-stationary training targets for both models, and, for DeepGMM, the adversarial training procedure of two models is similar to that of generative adversarial Networks (GANs), which are known to be unstable and difficult to train. In addition, when the CMR problem is weakly identifiable, as in the case of a weak instrument, the algorithms may converge to local minima that are far away from the true solution in the face of instabilities in the algorithm.

We conclude that solving the CMRs for an imitator policy can be sensitive to the choice of solver as well as to the choice of hyperparameters. In addition, some IV regression algorithms do not work well with high-dimension inputs. Our IV algorithm of choice, DML-IV, provides a robust base for the DML-IL algorithm that demonstrated good performance across all tasks and environments. This demonstrates the benefit of using double machine learning, which can debias two-stage estimators and provide good empirical and theoretical convergence.

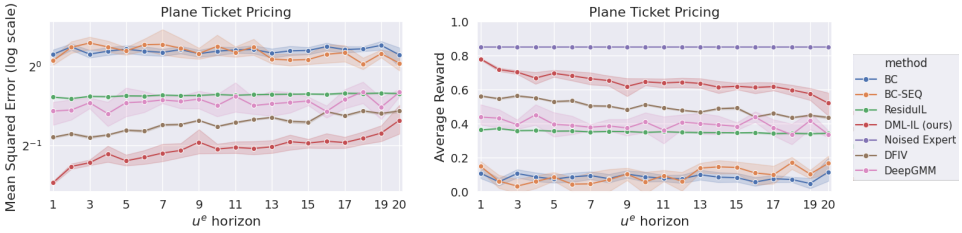


Figure 5: Additional results for the MSE between learnt policy and expert, and the average reward, in the airline ticket environment (Example 3.1), with DFIV and DeepGMM as the CMRs solver.

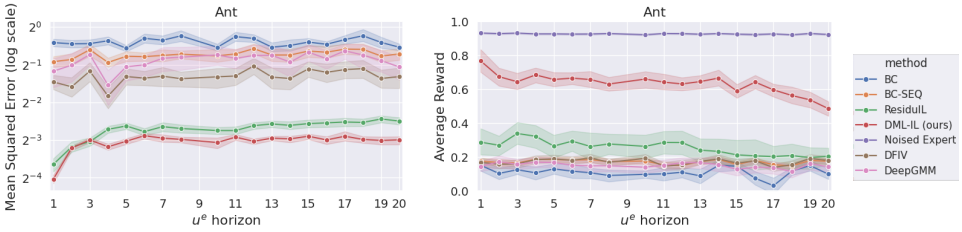


Figure 6: Additional results for the MSE between learnt policy and expert, and the average reward, Ant Mujoco environment, with DFIV and DeepGMM as the CMRs solver.

E Environments and Tasks

E.1 Dynamic airline Ticket Pricing

Here, we provide details regarding the dynamic airline ticket pricing environment introduced in Example 3.1. The environment and the expert policy are defined as follows:

$$\mathcal{S} := \mathbb{R} \tag{32}$$
$$\mathcal{A} := [-1, 1] \tag{33}$$
$$s_t = sign(s) \cdot u_t^o - u_t^\varepsilon \tag{34}$$
$$\pi_E = clip(-s/u_t^o, -1, 1) \tag{35}$$
$$a_t = \pi_E + 10 \cdot u_t^\varepsilon \tag{36}$$
$$u_t^o = mean(p_t \sim \text{Unif}[-1, 1], p_{t-1}, ...p_{t-M}) \tag{37}$$
$$u_t^\varepsilon = mean(q_t \sim \text{Normal}(0, 0.1 \cdot \sqrt{k}), q_{t-1}, ..., q_{t-k+1}) \tag{38}$$

where M is the influence horizon of the expert-observable u^o, which we set to 30. The states s_t are the profits at each time step, and the actions a_t are the final ticket price. u_t^o represent the seasonal patterns, where the expert π_E will try to adjust the price accordingly. u_t^ε represent the operating costs, which are additive both to the profit and price. Both u_t^o and u_t^ε are the mean over a set of i.i.d. samples, q_t and p_t, and vary across the time steps by updating the elements in the set at each time step. This construction allows u_t^ε and u_{t-k}^ε to be independent, since all set elements q_t will be resampled from time step $t-k$ to t. We multiply the standard deviation of q_t by $\sqrt{k}$ to make sure u_t^ε, which is the average over k i.i.d variables, have the same standard deviation for all choices of k.

E.2 MUJOCO ENVIRONMENTS

We evaluate DML-IL on three Mujoco environments: Ant, Half Cheetah and Hopper. The original tasks do not contain hidden variables, so we modify the environment to introduce u^{ε} and u^{o}. We use the default transition, state, and action space defined in the Mujoco environment. However, we changed the task objectives by altering the reward function and added confounding noise to both the state and the action. Specifically, instead of controlling the ant, half cheetah and hopper, respectively, to travel as fast as possible, the goal is to control the agent to travel at a target speed that is varying throughout an episode. This target speed is u^{o}, which is observed by the expert but not recorded in the dataset. In addition, we add confounding noise u_t^{ε} to s_t and a_t to mimic the environment noise such as wind noise. In all cases, the target speed u_t^{o}, confounding noise u_t^{ε} and the action a_t are generated as follows:

$$a_t = \pi_E + 20 \cdot u_t^{\varepsilon} \tag{39}$$

$$u_t^{o} = mean(p_t \sim \text{Unif}[-2, 4], p_{t-1}, ...p_{t-M}) \tag{40}$$

$$u_t^{\varepsilon} = mean(q_t \sim \text{Normal}(0, 0.01 \cdot \sqrt{k}), q_{t-1}, ..., q_{t-k+1}) \tag{41}$$

where $M = 30$, the state transitions follow the default Mujoco environment and the expert policy π_E is learned online in the environment. u_t^{o} and u_t^{ε} follow the airline ticket pricing environment to be the average over a queue of i.i.d. random variables. The reward is defined as $1_{healthy} -$ (current velocity $- u_t^{o})^2 -$ control loss, where $1_{healthy}$ gives reward 1 as long as the agent is in a healthy state as defined in the Mujoco documentation. The second penalty term penalises the deviation between the velocity of the current agent and the target velocity u_t^{o}. The control loss term is also as defined in default Mujoco, which is $0.1 * \sum(a_t^2)$ at each step to regularise the size of the actions.

E.2.1 ANT

In the Ant environment, we follow the gym implementation [3] with 8-dimensional action space and 28-dimensional observable state space, where the agent's position is also included in the state space. Since the target speed u_t^{o} is not recorded in the trajectory dataset, we scale the current position of the agent with respect to the target speed, $pos_t' = pos_{t-1} + \frac{pos_t - pos_{t-1}}{u_t^{o}}$, and use the new agent position pos_t' in the observed states. This allows the imitator to infer information regarding u_t^{o} from the trajectory history, namely from the rate of change in the past positions.

E.2.2 HALF CHEETAH

In the Half Cheetah environment, we follow the gym implementation [4] with 6-dimensional action space and 18-dimensional observable state space, where the agent's position is also included in the state space. Similarly to the Ant environment, we scale the current position of the agent to $pos_t' = pos_{t-1} + \frac{pos_t - pos_{t-1}}{u_t^{o}}$ such that the imitator can infer information regarding u_t^{o} from the trajectory history.

E.2.3 HOPPER

In the Hopper environment, we follow the gym implementation [5] with a 3-dimensional action space and a 12-dimensional observable state space, where the agent's position is also included in the state space. Similarly to the Ant environment, we scale the current position of the agent to $pos_t' = pos_{t-1} + \frac{pos_t - pos_{t-1}}{u_t^{o}}$ such that the imitator can infer information regarding u_t^{o} from trajectory history.

[3] Ant environment: `https://www.gymlibrary.dev/environments/mujoco/ant/`

[4] Half Cheetah environment: `https://www.gymlibrary.dev/environments/mujoco/half_cheetah/`

[5] Hopper environment: `https://www.gymlibrary.dev/environments/mujoco/hopper/`

Table 1: Network architecture for DML-IL. For mixture of Gaussians output, we report the number of components. No dropout is used.

(a) Roll-out model $\hat{M}$

Layer Type	**Configuration**
Input	state dim $\times$ 3
FC + ReLU	Out: 256
FC + ReLU	Out: 256
MixtureGaussian	5 components; Out: state dim $\times$ k

(b) Policy model $\hat{\pi}_h$

Layer Type	**Configuration**
Input	state dim$\times$ (k+3)
FC + ReLU	Out: 256
FC + ReLU	Out: 256
FC	Out: action dim

F Implementation Details

F.1 Expert Training

The expert in the airline ticket pricing environment is explicitly hand-crafted. For the Mujoco environments, we used the Stable-Baselines3 Raffin et al. (2021) implementation of soft actor-critic (SAC) and the default hyperparameters for each task outlined by Stable-Baseline3. The expert policy is an MLP with two hidden layers of size 256 and ReLU activations, and we train the expert for 10^7 steps.

F.2 Imitator Training

With the expert policy π_E, we generate 40 expert trajectories, each of 500 steps, following our previously defined environments. Specifically, the confounding noise is added to the state and actions and crucially u_t^o is not recorded in the trajectories. The naive BC directly learns $\mathbb{E}[a_t \mid s_t]$ through supervised learning. ResiduIL mainly follows the implementation of Swamy et al. (2022b), where we adopt it to allow a longer confounding horizon $k > 1$. For DML-IL and BC-SEQ, a history-dependent policy is used, where we fixed the look-back length to be $k + 3$, where k is the confounding horizon. BC-SEQ then just learns $\mathbb{E}[a_t \mid h_t]$ via supervised learning, and DML-IL is implemented with K-fold following Algorithm 2. The policy network architecture for BC, BC-SEQ and ResiduIL are 2 layer MLPs with 256 hidden size. The policy network $\hat{\pi}_h$ and the mixture of Gaussians roll-out model $\hat{M}$ for DML-IL have similar architecture, with details provided in Table 1. We use AdamW optimizer with weight decay of 10^{-4} and learning rate of 10^{-4}. The batch size is 64 and each model is trained for 150 epochs, which is sufficient for their convergence.

F.3 Imitator Evaluation

The trained imitator is then evaluated for 50 episodes, each 500 steps in the respective confounded environments. The average reward and the mean squared error between the imitator's action and the expert's action are recorded.

G Discussion regarding DML-IL

DML-IL can also be implemented with K-fold cross-fitting, where the dataset is partitioned into K folds, with each fold alternately used to train $\hat{\pi}_h$ and the remaining folds to train $\hat{M}$. This ensures unbiased estimation and improves the stability of training. The base IV algorithm DML-IV with K-fold cross-fitting is theoretically shown to converge at the rate of $O(N^{-1/2})$ Shao et al. (2024), where

Algorithm 2 DML-IL with K-fold cross-fitting

Input: Dataset $\mathcal{D}_E$ of expert demonstrations, Confounding noise horizon k, number of folds K for cross-fitting
Output: A history-dependent imitator policy $\hat{\pi}_h$
Get a partition $(I_k)_{k=1}^K$ of dataset indices $[N]$ of trajectories
for $k = 1$ **to** K **do**
  $I_k^c \coloneqq [N] \setminus I_k$
  Initialize the roll-out model $\hat{M}_i$ as a mixture of Gaussians model
  repeat
    Sample (h_t, a_t) from data $\{(\mathcal{D}_{E,i}) : i \in I_k^c\}$
    Fit the roll-out model $(h_t, a_t) \sim \hat{M}_i(h_{t-k})$ to maximize log likelihood
  until convergence
end for
Initialize the expert model $\hat{\pi}_h$ as a neural network
repeat
  for $k = 1$ **to** K **do**
    Sample h_{t-k} from $\{(\mathcal{D}_{E,i}) : i \in I_k\}$
    Generate $\hat{h}_t$ and $\hat{a}_t$ using the roll-out model $\hat{M}_i$
    Update $\hat{\pi}_h$ to minimise the loss $\ell := \|\hat{a}_t - \hat{\pi}_h(\hat{h}_t)\|_2$
  end for
until convergence

N is the sample size, under regularity conditions. DML-IL with K-fold cross-fitting (see Appendix G for details) will thus inherit this convergence rate guarantee.

Note that Algorithm 1 requires the confounding noise horizon k as input. Although the exact value of k can be difficult to obtain in reality, any upper bound $\bar{k}$ of k is sufficient to guarantee the correctness of Algorithm 1, since $h_{t-\bar{k}}$ is also a valid instrument. Ideally, we would like a data-driven approach to determine k. Unfortunately, it is generally intractable to empirically verify whether h_{t-k} is a valid instrument from a static dataset, especially the unconfounded instrument condition (i.e., $h_{t-k} \perp\!\!\!\perp u_t^\varepsilon$). Therefore, we rely on the user to provide a sensible choice of $\bar{k}$ based on the environment that does not substantially overestimate k.

G.1 DML-IL WITH K-FOLD CROSS-FITTING

Here, we outline DML-IL with K-fold cross-fitting, which ensures unbiased estimation and improves training stability. The algorithm is shown in Algorithm 2. The dataset is partitioned into K folds based on the trajectory index. For each fold, we use the leave-out data, that is, indices $I_k^c \coloneqq [N] \setminus I_k$, to train separate roll-out models $\hat{M}_i$ for $i \in [1..K]$. Then, to train a single expert model $\hat{\pi}_h$, we sample the trajectory history h_{t-k} from each fold and use the roll-out model trained with the leave-out data to complete the trajectory and train $\hat{\pi}_h$. This technique is very important in Double Machine Learning (DML) literature Shao et al. (2024), Chernozhukov et al. (2018) for it provides both empirical stability and theoretical guarantees. The base IV regression algorithm DML-IV with K-fold cross-fitting is theoretically shown to converge at the rate of $O(N^{-1/2})$ Shao et al. (2024), where N is the sample size, under technical regularity and identifiability conditions (see Shao et al. (2024) for the technical conditions). These conditions are typically assumed for similar theoretical analyses, and DML-IL with K-fold cross-fitting will thus inherit this convergence rate guarantee if the regularity conditions are satisfied.