CAUSAL IMITATION LEARNING VIA INVERSE REINFORCEMENT LEARNING

Kangrui Ruan∗, Junzhe Zhang∗, Xuan Di, and Elias Bareinboim
Columbia University, New York, NY 10027, USA
{kr2910,junzhez,sharon.di,eliasb}@columbia.edu

ABSTRACT

One of the most common ways children learn when unfamiliar with the environment is by mimicking adults. Imitation learning concerns an imitator learning to behave in an unknown environment from an expert’s demonstration; reward signals remain latent to the imitator. This paper studies imitation learning through causal lenses and extends the analysis and tools developed for behavior cloning (Zhang, Kumor, Bareinboim, 2020) to inverse reinforcement learning. First, we propose novel graphical conditions that allow the imitator to learn a policy performing as well as the expert’s behavior policy, even when the imitator and the expert’s state-action space disagree, and unobserved confounders (UCs) are present. When provided with parametric knowledge about the unknown reward function, such a policy may outperform the expert’s. Also, our method is easily extensible and allows one to leverage existing IRL algorithms even when UCs are present, including the multiplicative-weights algorithm (MWAL) (Syed & Schapire, 2008) and the generative adversarial imitation learning (GAIL) (Ho & Ermon, 2016). Finally, we validate our framework by simulations using real-world and synthetic data.

1 INTRODUCTION

Reinforcement Learning (RL) has been deployed and shown to perform extremely well in highly complex environments in the past decades (Sutton & Barto, 1998; Mnih et al., 2013; Silver et al., 2016; Berner et al., 2019). One of the critical assumptions behind many of the classical RL algorithms is that the reward signal is fully observed, and the reward function could be well-specified. In many real-world applications, however, it might be impractical to design a suitable reward function that evaluates each and every scenario (Randlov & Alstrøm, 1998; Ng et al., 1999). For example, in the context of human driving, it is challenging to design a precise reward function, and experimenting in the environment could be ill-advised; still, watching expert drivers operating is usually feasible.

In machine learning, the imitation learning paradigm investigates the problem of how an agent should behave and learn in an environment with an unknown reward function by observing demonstrations from a human expert (Argall et al., 2009; Billard et al., 2008; Hussein et al., 2017; Osa et al., 2018). There are two major learning modalities that implements IL – behavioral cloning (BC) (Widrow, 1964; Pomerleau, 1989; Muller et al., 2006; Mulling et al., 2013; Mahler & Goldberg, 2017) and inverse reinforcement learning (IRL) (Ng et al., 2000; Ziebart et al., 2008; Ho & Ermon, 2016; Fu et al., 2017). BC methods directly mimic the expert’s behavior policy by learning a mapping from observed states to the expert’s action via supervised learning. Alternatively, IRL methods first learn a potential reward function under which the expert’s behavior policy is optimal. The imitator then obtains a policy by employing standard RL methods to maximize the learned reward function. Under some common assumptions, both BC and IRL are able to obtain policies that achieve the expert’s performance (Kumor et al., 2021; Swamy et al., 2021). Moreover, when additional parametric knowledge about the reward function is provided, IRL may produce a policy that outperforms the expert’s in the underlying environment (Syed & Schapire, 2008; Li et al., 2017; Yu et al., 2020).

For concreteness, consider a learning scenario depicted in Fig. 1a, describing trajectories of human-driven cars collected by drones flying over highways (Krajewski et al., 2018; Etesami & Geiger, 2020). Using such data, we want to learn a policy $X \leftarrow \pi(Z)$ deciding on the acceleration (action) $X \in \mathbb{R}$.
the expected reward 

\[
E[Y] = P(X = 1, Z = 0) + P(X = 0, Z = 1) = 0.5
\]

Now we apply standard IRL algorithms to learn a policy \( X \leftarrow \pi(Z) \) so that the imitator’s driving performance, denoted by \( E[Y \mid \text{do}(\pi)] \), is at least as good as the expert’s performance \( E[Y] \). Detailed derivations of IRL policy are shown in Ruan et al. (2023) Appendix A. Note that \( E[Y \mid z, x] = x + z - 2xz \) belongs to a family of reward functions \( f_Y(x, z) = \alpha x + \beta z - \gamma xz \), where \( 0 < \alpha < \gamma \).

A typical IRL imitator solves a minimax problem

\[
\min_{\pi} \max_{\pi} E[f_Y(X, Z)] - E[f_Y(X, Z) \mid \text{do}(\pi)].
\]

The inner step “guesses” a reward function being optimized by the expert; while the outer step learns a policy maximizing the learned reward function. Applying these steps leads to a policy \( \pi^* : X \leftarrow \neg Z \) with the expected reward \( E[Y \mid \text{do}(\pi^*)] = 1 \), which outperforms the sub-optimal expert.

Despite the performance guarantees provided by existing imitation methods, both BC and IRL rely on the assumption that the expert’s input observations match those available to the imitator. More recently, there exists an emerging line of research under the rubric of causal imitation learning that augments the imitation paradigm to account for environments consisting of arbitrary causal mechanisms and the aforementioned mismatch between expert and imitator’s sensory capabilities (de Haan et al., 2019; Zhang et al., 2020; Etesami & Geiger, 2020; Kumor et al., 2021). Closest to our work, Zhang et al. (2020); Kumor et al. (2021) derived graphical criteria that completely characterize when and how BC could lead to successful imitation even when the agents perceive reality differently. Still, it is unclear how to perform IRL-type training if some expert’s observed states remain latent to the imitator, which leads to the presence of unobserved confounding (UCs) in expert’s demonstrations. Perhaps surprisingly, naïvely applying IRL methods when UCs are present does not necessarily lead to satisfactory performance, even when the expert itself behaves optimally.

To witness, we now modify the previous highway driving scenario to demonstrate the challenges of UCs. In reality, covariates \( Z \) (i.e., velocities and location) are also affected by the car horn \( U_1 \) of surrounding vehicles and the wind condition \( U_2 \). However, due to the different perspectives of drones (recording from the top), such critical information (i.e., \( U_1, U_2 \)) is not recorded by the camera and thus remains unobserved. Fig. [15] graphically describes this modified learning setting. More specifically, consider an instance where \( Z \leftarrow U_1 \oplus U_2, Y \leftarrow X \oplus Z \oplus U_2; \oplus \) is the exclusive-or operator; and values of \( U_1 \) and \( U_2 \) are drawn uniformly over \( \{0, 1\} \). An expert driver, being able to hear the car horn \( U_1 \), follows a behavior policy \( X \leftarrow U_1 \) and achieves the optimal performance \( E[Y] = 1 \).

Meanwhile, observe that \( E[Y \mid z, x] = 1 \) belongs to a family of reward functions \( f_Y(x, z) = \alpha \) (where \( \alpha > 0 \)). Solving \( \min_{\pi} \max_{\pi} E[f_Y(X, Z)] - E[f_Y(X, Z) \mid \text{do}(\pi)] \) leads to an IRL policy \( \pi^* \) with expected reward \( E[Y \mid \text{do}(\pi^*)] = 0.5 \), which is far from the expert’s optimal performance \( E[Y] = 1 \).

After all, a question that naturally arises is, under what conditions an IRL imitator procedure can perform well when UCs are present, and there is a mismatch between the perception of the two agents? In this paper, we answer this question and, more broadly, investigate the challenge of performing IRL through causal lenses. In particular, our contributions are summarized as follows. (1) We provide a novel, causal formulation of the inverse reinforcement learning problem. This formulation allows one to formally study and understand the conditions under which an IRL policy is learnable, including in settings where UCs cannot be ruled out a priori. (2) We derive a new graphical condition for deciding whether an imitating policy can be computed from the available data and knowledge, which provides a robust generalization of current IRL algorithms to non-Markovian settings, including GAIL (Ho & Ermon, 2016) and MWAL (Syed & Schapire, 2008). (3) Finally, we move beyond this graphical condition and develop an effective IRL algorithm for structural causal models (Pearl, 2000) with
We investigate the sequential decision-making setting concerning a set of actions. The basic semantic framework of our analysis rests on structural causal models (SCMs) (Chakraborty & Murphy, 2014), which generalize personalized medicine to time-varying treatment policies (Murphy et al., 2001; Chakraborty & Murphy, 2014). Such policies π over actions X is a sequence of decision rules π = {π₁, . . . , πⁿ}. Each decision rule πᵢ(Xᵢ | Zᵢ) is a probability distribution over an action Xᵢ ∈ X, conditioning on values of a set of covariates Zᵢ ⊆ PAᵢ. Such policies π are also referred to as dynamic treatment regimes (Murphy et al., 2001; Bareinboim et al., 2022), which generalize personalized medicine to time-varying treatment settings in healthcare, in which treatment is repeatedly tailored to a patient’s dynamic state.

A policy intervention on actions X following a policy π, denoted by do(π), entails a submodel Mᵢ from a SCM M where structural functions f_X associated with X (i.e., the expert’s behavior policy) are replaced with decision rules Xᵢ ∼ πᵢ(Xᵢ | Zᵢ) for every Xᵢ ∈ X. A critical assumption on actions X and small letters for their values (x). Dₓ represents the domain of X and Dₓ the space of probability distributions over Dₓ. For a set X, let |X| denote its dimension. The probability distribution over variables X is denoted by P(X).

Similarly, P(Y | X) represents a set of conditional distributions P(Y | X = x) for all realizations x. We use abbreviations P(x) for probabilities P(X = x); so does P(Y = y | X = x) = P(y | x). Finally, indicator function 1{Z = z} returns 1 if Z = z holds true; otherwise 0.

The basic semantic framework of our analysis rests on structural causal models (SCMs) (Pearl 2000 Ch. 7). An SCM M is a tuple (U, V, F, P(U)) with V the set of endogenous, and U exogenous variables. F is a set of structural functions s.t. for fᵥ ∈ F, V ← fᵥ(paᵥ, uᵥ), with PAᵥ ⊆ V ∪ U. Values of U are drawn from an exogenous distribution P(U), inducing distribution P(V) over endogenous variables V. Since the learner can observe only a subset of endogenous variables, we split V into a partition O ∪ L where variable O ⊆ V are observed and L = V \ O remain latent to the learner. The marginal distribution P(O) is thus referred to as the observational distribution. An atomic intervention on a subset X ⊆ V, denoted by do(x), is an operation where values of X are set to constants x, replacing the functions fₓ = {fₓ: ∀X ∈ X} that would normally determine their values. For an SCM M, let Mᵢ be a submodel of M induced by intervention do(x). For a set Y ⊆ V, the interventional distribution P(s(do(x)) induced by do(x) is defined as the distribution over Y in the submodel Mᵢ, i.e., P_Mᵢ(Y).

2 Causal Inverse Reinforcement Learning

We investigate the sequential decision-making setting concerning a set of actions X, a series of covariates Z, and a latent reward Y in an SCM M. An expert (e.g., a physician, driver), operating in SCM M, selects actions following a behavior policy, which is the collection of structural functions f_X = {f_X | X ∈ X}. The expert’s performance is evaluated as the expected reward E[Y]. On the other hand, a learning agent (i.e., the imitator) intervenes on actions X following an ordering X₁ ⊲ ··· ⊲ Xₙ; each action Xᵢ is associated with a set of features PAᵢ ⊆ O \ {Xᵢ}. A policy π over actions X is a sequence of decision rules π = {π₁, . . . , πᵢ}. Each decision rule πᵢ(Xᵢ | Zᵢ) is a probability distribution over an action Xᵢ ∈ X, conditioning on values of a set of covariates Zᵢ ⊆ PAᵢ. Such policies π are also referred to as dynamic treatment regimes (Murphy et al., 2001; Chakraborty & Murphy, 2014), which generalize personalized medicine to time-varying treatment settings in healthcare, in which treatment is repeatedly tailored to a patient’s dynamic state.

A policy intervention on actions X following a policy π, denoted by do(π), entails a submodel Mᵢ from a SCM M where structural functions f_X associated with X (i.e., the expert’s behavior policy) are replaced with decision rules Xᵢ ∼ πᵢ(Xᵢ | Zᵢ) for every Xᵢ ∈ X. A critical assumption...
Throughout this paper, detailed parametrizations of the underlying SCM $M$ are assumed to be unknown to the agent. Instead, the agent has access to the input: (1) a causal diagram $G$ associated with $M$, and (2) the expert’s demonstrations, summarized as the observational distribution $P(O)$. The goal of the agent is to output an imitating policy $\pi^*$ that achieves the expert’s performance.

**Definition 1.** For an SCM $M = (U, V, F, P(U))$, an imitating policy $\pi^*$ is a policy such that its expected reward is lower bounded by the expert’s reward, i.e., $\mathbb{E}_M[Y \mid \text{do}(\pi^*)] \geq \mathbb{E}_M[Y]$.

In words, the right-hand side is the expert’s performance that the agent wants to achieve, while the left-hand side is the real reward experienced by the agent. The challenge in imitation learning arises from the fact that the reward $Y$ is not specified and latent, i.e., $Y \not\in O$. This precludes approaches that identify $\mathbb{E}[Y \mid \text{do}(\pi)]$ directly from the demonstration data (e.g., through the do- or soft-do-calculus by Pearl (2000). Correa & Bareinboim (2020)).

There exist methods in the literature for finding an imitating policy in Def 1. Before describing their details, we first introduce some necessary concepts. For any policy $\pi$, we summarize its associated state-action domain using a sequence of pairs of variables called a policy scope $S$.

**Definition 2.** (Lee & Bareinboim (2020)). For an SCM $M$, a policy scope $S$ (short, scope) over actions $X$ is a sequence of tuples $\{(X_i, Z_i)\}_{i=1}^n$ where $Z_i \subseteq PA_i$ for every $X_i \in X$.

We will consistently use $\pi \sim S$ to denote a policy $\pi$ associated with scope $S$. For example, consider a policy scope $S = \{(X_1, Z_1), (X_2, Z_2)\}$ over actions $X_1, X_2$ in Fig. 1c. A policy $\pi \sim S$ is a sequence of distributions $\pi = \{(\pi_1(X_1 \mid Z_1), \pi_2(X_2 \mid Z_2))\}$.

(Zhang et al. (2020)). Given a causal diagram $G$, a policy scope $S = \{(X_i, Z_i)\}_{i=1}^n$ is said to satisfy the sequential $\pi$-backdoor criterion in $G$ (for short, $\pi$-backdoor admissible) if at each $X_i \in X$, one of the following conditions hold: (1) $X_i$ is not an ancestor of $Y$ in $G^{(i)}$, i.e., $X_i \notin AN(Y)_{G^{(i)}}$; or (2) $Z_i$ blocks all backdoor path from $X_i$ to $Y$ in $G^{(i)}$, i.e., $(Y \perp X_i | Z_i)_{G^{(i)}}$.

(Kumor et al. (2021)). A causal diagram $G$, a policy scope $S = \{(X_i, Z_i)\}_{i=1}^n$ is said to satisfy the sequential $\pi$-backdoor criterion in $G$ (for short, $\pi$-backdoor admissible) if at each $X_i \in X$, one of the following conditions hold: (1) $X_i$ is not an ancestor of $Y$ in $G^{(i)}$, i.e., $X_i \notin AN(Y)_{G^{(i)}}$; or (2) $Z_i$ blocks all backdoor path from $X_i$ to $Y$ in $G^{(i)}$, i.e., $(Y \perp X_i | Z_i)_{G^{(i)}}$.

2.1 **Minimal Sequential Backdoor Criterion**

To circumvent this issue, we take a somewhat different approach to causal imitation by incorporating the principle of inverse reinforcement learning (IRL) principle. Following the game-theoretic approach (Syed & Schapire (2008)), we formulate the problem as learning to play a two-player zero-sum game in which the agent chooses a policy, and the nature chooses an SCM instance. A key property of this algorithm is that it allows us to incorporate prior parametric knowledge about the latent reward signal. When such knowledge is informative, our algorithm is about to obtain a policy that could significantly outperform the expert with respect to the unknown causal environment, while at the same time are guaranteed to be no worse. Formally, let $\mathcal{M} = \{M \mid G = G, P_M(O) = P(O)\}$ denote the set of SCMs compatible with both the causal diagram $G$ and the observational distribution $P(O)$. Fix a policy scope $S$. Now consider the optimization problem defined as follows.

$$\nu^* = \min_{\pi \sim S} \max_{M \in \mathcal{M}} \mathbb{E}_M[Y] - \mathbb{E}_{M}[Y \mid \text{do}(\pi)].$$ (1)
The inner maximization in the above equation can be viewed as a causal IRL step where we attempt to "guess" a worst-case SCM $\hat{M}$ above the expert’s and the imitator’s policies is maximized. Meanwhile, since the expert’s reward $\mathbb{E}_M[Y]$ is not affected by the imitator’s policy $\pi$, the outer minimization is equivalent to a planning step that finds a policy $\pi^*$ optimizing the learned SCM $\hat{M}$. Obviously, the solution $\pi^*$ is an imitating policy if $\nu^* = 0$. In cases where the expert is sub-optimal, i.e., $E_{\pi^*}(Y) < E_{\pi}(Y | \|do(\pi))$ for some policies $\pi$, we may have $\nu^* < 0$. That is, the policy $\pi^*$ will dominate the expert’s policy $f_X$ regardless of parametrizations of SCM $M$ in the worst-case scenario. In other words, $\pi^*$ to some extent ignores the sub-optimal expert, and instead exploits prior knowledge about the underlying model.

Despite the clear semantics in terms of causal models, the optimization problem in Eq. (1) requires the learner to search over all possible SCMs compatible with the causal diagram $\mathcal{G}$ and observational distribution $P(O)$. In principle, it entails a quite challenging search since one does not have access to the parametric forms of the underlying structural functions $\mathcal{F}$ nor the exogenous distribution $P(U)$. It is not clear how the existing optimization procedures can be used.

In this paper, we will develop novel methods to circumvent this issue, thus leading to effective imitating policies. Our first algorithm relies on a refinement of the sequential $\pi$-backdoor, based on the concept of minimality. A subscope $S'$ of a policy scope $S = \{(X_i, Z_i)\}_{i=1}^n$, denoted by $S' \subseteq S$, is a sequence $\{(X_i, Z'_i)\}_{i=1}^n$ where $Z'_i \subseteq Z_i$ for every $X_i \in X$. A proper subscope $S' \subset S$ is a subscope in $S$ other than $S$ itself. The minimal $\pi$-backdoor admissible scope is defined as follows.

**Definition 4.** Given a causal diagram $\mathcal{G}$, a $\pi$-backdoor admissible scope $S$ is said to be minimal if there exists no proper subscope $S' \subset S$ satisfying the sequential $\pi$-backdoor in $\mathcal{G}$.

**Theorem 1.** Given a causal diagram $\mathcal{G}$, if there exists a minimal $\pi$-backdoor admissible scope $S = \{(X_i, Z_i)\}_{i=1}^n$ in $\mathcal{G}$, consider the following conditions:

1. Let effective actions $X^* = X \cap An(Y)_{\mathcal{G}_s}$ and effective covariates $Z^* = \bigcup_{X_i \in X^*} Z_i$;
2. For $i = 1, \ldots, n$, let $X^*_i = \{X_j \in X^* | j < i\}$ and $Z^*_{i} = \bigcup_{X_j \in X^*} Z_j$.

Then, for any policy $\pi \sim S$, the expected reward $\mathbb{E}[Y | do(\pi)]$ is computable from $P(O,Y)$ as:

$$\mathbb{E}[Y | do(\pi)] = \sum_{x^*, z^*} \mathbb{E}[Y | x^*, z^*] \rho_\pi(x^*, z^*)$$

(2)

where the occupancy measure $\rho_\pi(x^*, z^*) = \prod_{X_i \in X^*} P(z_i | x^*_i, z^*_{<i})$ $\pi_i(x_i | z_i)$.

To illustrate, consider again the causal diagram $\mathcal{G}$ in Fig. 1c, the manipulated diagram $\hat{G}^{(2)} = \mathcal{G}$ and $\hat{G}^{(1)}$ is obtained from $\mathcal{G}$ by removing $Z_2 \leftarrow X_2$. While scope $S_1 = \{(X_1, \{Z_1\}), \{X_2, \{Z_2\}\})$ satisfies the sequential $\pi$-backdoor, it is not minimal since $(X_1 \perp \perp Y)$ in $\hat{G}^{(1)}_{X_1}$. On the other hand, $S_2 = \{(X_1, \emptyset), \{X_2, \{Z_2\}\})$ is minimal $\pi$-backdoor admissible since $(X_2 \perp \perp Y | Z_2)$ holds true in $\hat{G}^{(2)}_{X_2}$; and the covariate set $\{Z_2\}$ is minimal due to the presence of the backdoor path $X_2 \leftarrow Z_2 \rightarrow Y$.

Let us focus on the minimal $\pi$-backdoor admissible scope $S_2$. Note that $G_{S_2}$ is a subgraph obtained from $G$ by removing the bi-directed arrow $Z_2 \leftrightarrow X_2$. We must have effective actions $X^* = \{X_1, X_2\}$ and effective covariates $Z^* = \{Z_2\}$. Therefore, $Z^*_{<1} = Z^*_{<2} = \emptyset$ and $Z^*_{<3} = \{Z_2\}$. For any policy $\pi \sim S_2$, Thm. 1 implies $\mathbb{E}[Y | do(\pi)] = \sum_{x_{1,2,3}} \mathbb{E}[Y | x_{1,2,3}] P(z_2 | x_1) \pi_2(x_2 | z_2) \pi_1(x_1)$. On the other hand, the same result in Thm. 1 does not necessarily hold for a non-minimal $\pi$-backdoor admissible scope. For instance, consider again the non-minimal scope $S_1 = \{(X_1, \{Z_1\}), \{X_2, \{Z_2\}\})$. The expected reward $\mathbb{E}[Y | do(\pi)]$ of a policy $\pi \sim S_2$ is not computable from Eq. (2), and is ultimately not identifiable from distribution $P(O,Y)$ in $\mathcal{G}$ (Tian, 2008).

2.2 Imitation via Inverse Reinforcement Learning

Once a minimal $\pi$-backdoor admissible scope $S$ is found, there exist effective procedures to solve for an imitating policy in Eq. (1). Let $\mathcal{A}$ be a hypothesis class containing all expected rewards $\mathbb{E}_M[Y | x^*, z^*] \subset \mathcal{M} \subset \mathcal{A}$, i.e., $\mathcal{A} = \{\mathbb{E}_M[Y | x^*, z^*] | \forall M \in \mathcal{M}\}$. Applying the identification formula in Thm. 1 reduces the optimization problem in Eq. (1) as follows:

$$\nu^* = \min_{\pi \sim S} \max_{r \in \mathcal{A}} \sum_{x^*, z^*} r(x^*, z^*) (\rho_\pi(x^*, z^*) - \rho_\pi(x^*, z^*))$$

(3)
where the expert’s occupancy measure \( \rho(x^*, z^*) = P(x^*, z^*) \) and the agent’s occupancy measure \( \rho_\pi(x^*, z^*) \) is given by Eq. (3). The above minimax problem is solvable using standard IRL algorithms. The identification result in Thm. 1 ensures that the learned policy applies to any SCM compatible with the causal diagram and the observational data, thus robust to the unobserved confounding bias in the expert’s demonstrations. Henceforth, we will consistently refer to Eq. (3) as the \textit{canonical equation of causal IRL}. In this paper, we solve for an imitating policy \( \pi^* \) in Eq. (3) using state-of-the-art IRL algorithms, provided with common choices of parametric reward functions. These algorithms include the multiplicative-weights algorithm (MWAL) \cite{Syed2008} and the generative adversarial imitation learning (GAIL) \cite{Ho2016}. We refer readers to Algs. 3 and 4 in \cite{Ruan2023} Appendix C for more discussions on the pseudo-code and implementation details.

\textbf{Causal MWAL} \cite{Abbeel2004, Syed2008} study IRL in Markov decision processes where the reward function \( r(x^*, z^*) \) is a linear combination of \( k \)-length feature expectations vectors \( \phi(x^*, z^*) \). Particularly, let \( r(x^*, z^*) = \mathbf{w} \cdot \phi(x^*, z^*) \) for a coefficient vector \( \mathbf{w} \) contained in a convex set \( \mathbf{S}^k = \{ \mathbf{w} \in \mathbb{R}^k \mid \|\mathbf{w}\|_1 = 1 \text{ and } \mathbf{w} \geq 0 \} \). Let \( \phi^{(i)} \) be the \( i \)-th component of feature vector \( \phi \) and let deterministic policies with scope \( \mathcal{S} \) be ordered by \( \pi^{(1)}, \ldots, \pi^{(n)} \). The canonical equation in Eq. (3) is reducible to a two-person zero-sum matrix game under linearity.

\textbf{Proposition 1.} For a hypothesis class \( \mathcal{R} = \{ r = \mathbf{w} \cdot \phi \mid \mathbf{w} \in \mathbf{S}^k \} \), the solution \( \nu^* \) of the canonical equation in Eq. (3) is obtainable by solving the following minimax problem:

\[
\nu^* = \min_{\pi \sim \mathcal{R}} \max_{\mathbf{w} \in \mathbf{S}^k} \mathbf{G} \pi,
\]

where \( \mathbf{G} \) is a \( k \times n \) matrix given by \( G(i, j) = \sum_{x^*, z^*} \phi^{(i)}(x^*, z^*) (\rho(x^*, z^*) - \rho_{\pi^{(j)}}(x^*, z^*)) \).

There exist effective multiplicative weights algorithms for solving the matrix game in Eq. (4), including MW (Freund & Schapire 1999) and MW AL (Syed & Schapire 2008).

\textbf{Causal GAIL} \cite{Ho2016} introduces the GAIL algorithm for learning an imitating policy in Markov decision processes with a general family of non-linear reward functions. In particular, \( r(x^*, z^*) \) takes values in the real space \( \mathbb{R} \), i.e., \( r \in \mathbb{R}^{X^* \times Z^*} \) where \( \mathbb{R}^{X^* \times Z^*} = \{ r : \mathcal{D}_{X^*} \times \mathcal{D}_{Z^*} \mapsto \mathbb{R} \} \). The complexity of reward function \( r \) is penalized by a vector regularization function \( \psi(r) \), i.e.,

\[
\nu^* = \min_{\pi \sim \mathcal{R}} \max_{r \in \mathbb{R}^{X^* \times Z^*}} \sum_{x^*, z^*} r(x^*, z^*) (\rho(x^*, z^*) - \rho_{\pi}(x^*, z^*)) - \psi(r)
\]

Henceforth, we will consistently refer to Eq. (5) as the \textit{penalized canonical equation of causal IRL}. It is often preferable to solve its conjugate form. Formally,

\textbf{Proposition 2.} For a hypothesis class \( \mathcal{R} = \{ r : \mathcal{D}_{X^*} \times \mathcal{D}_{Z^*} \mapsto \mathbb{R} \} \) regularized by \( \psi \), the solution \( \nu^* \) of the penalized canonical equation in Eq. (5) is obtainable by solving the following problem:

\[
\nu^* = \min_{\pi \sim \mathcal{S}} \psi^*(\rho - \rho_{\pi})
\]

where \( \psi^* \) be a conjugate function of \( \psi \) and is given by \( \psi^* = \max_{r \in \mathbb{R}^{X^* \times Z^*}} a^\top r - \psi(r) \).

Eq. (6) seeks a policy \( \pi \) which minimizes the divergence of the occupancy measures between the imitator and the expert, as measured by the function \( \psi^* \). The computational framework of generative adversarial networks \cite{Goodfellow2014} provides an effective approach to solve such a matching problem, e.g., the GAIL algorithm \cite{Ho2016}.

\section{Causal Imitation without Sequential Backdoor}

In this section, we investigate causal IRL beyond the condition of minimal sequential \( \pi \)-backdoor. Observe that the key to the reduction of the canonical causal IRL equation in Eq. (3) lies in the identification of expected rewards \( \mathbb{E}[Y \mid \text{do}(\pi)] \) had the latent reward \( Y \) been observed. Next we will study general conditions under which \( \mathbb{E}[Y \mid \text{do}(\pi)] \) is uniquely discernible from distribution \( P(O, Y) \) in the causal diagram \( \mathcal{G} \), called the \textit{identifiability} of causal effects \cite{Pearl2000} Def. 3.2.4.

\textbf{Definition 5 (Identifiability).} Given a causal diagram \( \mathcal{G} \) and a policy \( \pi \sim \mathcal{S} \), the expected reward \( \mathbb{E}[Y \mid \text{do}(\pi)] \) is said to be identifiable from distribution \( P(O, Y) \) in \( \mathcal{G} \) if \( \mathbb{E}[Y \mid \text{do}(\pi)] \) is uniquely computable from \( P(O, Y) \) in any SCM \( M \) compatible with \( \mathcal{G} \).
We say a policy scope $S$ is identifiable (from $P(O,Y)$ in $\mathcal{G}$) if for all policies $\pi \sim S$, the corresponding expected rewards $E[Y \mid do(\pi)]$ are identifiable from $P(O,Y)$ in $\mathcal{G}$. Our next result shows that whenever an identifiable policy scope $S$ is found, one could always reduce the causal IRL problem to the canonical optimization equation in Eq. (3).

**Theorem 2.** Given a causal diagram $\mathcal{G}$, a policy scope $S$ is identifiable from $P(O,Y)$ in $\mathcal{G}$ if and only if for any policy $\pi \sim S$, the expected reward $E[Y \mid do(\pi)]$ is computable from $P(O,Y)$ as

$$E[Y \mid do(\pi)] = \sum_{x^*,z^*} E[Y \mid x^*, z^*] \rho_\pi(x^*, z^*)$$

where subsets $X^* \subseteq X$, $Z^* \subseteq O \setminus X$; and the imitator’s occupancy measure $\rho_\pi(x^*, z^*)$ is a function of the observational distribution $P(O)$ and policy $\pi$.

Thm. 2 suggests a general procedure to learn an imitating policy via causal IRL. Whenever an identifiable scope $S$ is found, the identification formula in Eq. (7) permits one to reduce the optimization problem in Eq. (1) to the canonical equation in Eq. (3). One could thus obtain an imitating policy $\pi \sim S$ by solving Eq. (3) where the expert’s occupancy measure $\rho(x^*, z^*) = P(x^*, z^*)$ and the imitator’s occupancy measure $\rho_\pi(x^*, z^*)$ is given by Eq. (7). As an example, consider the frontdoor diagram described in Fig. 2a and a policy scope $S = \{X, \emptyset\}$. The expected reward $E[Y \mid do(\pi)] = \sum_{x'} E[Y \mid do(x')] \pi(x')$ and $E[Y \mid do(x')]$ is identifiable from $P(X,Y,Z)$ using the frontdoor adjustment formula (Pearl 2000 Thm. 3.3.4). The expected reward $E[Y \mid do(\pi)]$ of any policy $\pi(X)$ could be written as:

$$E[Y \mid do(\pi)] = \sum_{x,z} E[Y \mid x, z] P(x) \sum_{x'} P(z \mid x') \pi(x').$$

Let occupancy measures $\rho(x, z) = P(x, z)$ and $\rho_\pi(x, z) = P(x) \sum_{x'} P(z \mid x') \pi(x')$. We could thus learn an imitating policy in the frontdoor diagram by solving the canonical equation given by:

$$\nu^* = \min_{\pi \sim S} \max_{r \in \mathcal{R}} \sum_{x,z} r(x,z) (\rho(x,z) - \rho_\pi(x,z)).$$

where $\mathcal{R}$ is a hypothesis class of the reward function $r(x,z) \triangleq E[Y \mid x, z]$. The solution $\pi^*(X)$ is an imitating policy performing at least as well as the expert’s behavior policy if the gap $\nu^* \leq 0$.

Next, we will describe how to obtain the identification formula in Eq. (7) provided with an identifiable scope $S$. Without loss of generality, we will assume that the reward $Y$ is the only endogenous variable that is latent in the causal diagram $\mathcal{G}$, i.e., $V = O \cup \{Y\}$. We will utilize a special type of clustering of nodes in the causal diagram $\mathcal{G}$, called the *confounded component* (for short, c-component).

**Definition 6** (C-component (Tian & Pearl 2002)). For a causal diagram $\mathcal{G}$, a subset $C \subseteq V$ is a c-component if any pair $V_i, V_j \in C$ is connected by a bi-directed path in $\mathcal{G}$.

For instance, the frontdoor diagram in Fig. 2a contains two c-components $C_1 = \{X,Y\}$ and $C_2 = \{Z\}$. We will use a sound and complete procedure Identify (Tian 2002; 2008) for identifying causal effects $E[Y \mid do(\pi)]$ of an arbitrary policy $\pi \sim S$. Particularly, Identify takes as input the causal diagram $\mathcal{G}$, a reward $Y$, and a policy scope $S$. It returns an identification formula for $E[Y \mid do(\pi)]$ from $P(O,Y)$ if expected rewards of all policies $\pi \sim S$ are identifiable. Otherwise, Identify($\mathcal{G}, Y, S$) = “Fail”. Details of Identify are shown in (Zhang et al. 2020 Appendix B). Recall that $G_S$ is the causal diagram of submodel $M_\pi$ induced by policy $\pi \sim S$. Fig. 2b shows diagram $G_S$ obtained from the frontdoor graph $\mathcal{G}$ and scope $S = \{(X, \emptyset)\}$ as described in Fig. 2a. Let $Z_\mathcal{G} = An(Y)$ be ancestors of $Y$ in $G_\mathcal{G}$. Our next result states that Identify($\mathcal{G}, Y, S$) is ensured to find an identification formula of the form in Eq. (7) when it is identifiable.

**Lemma 1.** Given a causal diagram $\mathcal{G}$, a policy scope $S$ is identifiable from $P(O,Y)$ in $\mathcal{G}$ if and only if Identify($\mathcal{G}, Y, S$) ≠ “Fail”. Moreover, Identify($\mathcal{G}, Y, S$) returns an identification formula of the form in Eq. (7) where $X^* = Pa(C_Y) \cap X$ and $Z^* = Pa(C_Y) \setminus \{\{Y\} \cup X\}$; and $C_Y$ is a c-component containing reward $Y$ in subgraph $G_{\{An(Z_Y)\}}$.

Otherwise, one could always simplify the diagram $\mathcal{G}$ and project other latent variables $L \setminus \{Y\}$ using the projection algorithm (Tian 2002 Sec. 4.5), without affecting the identifiability of target query $E[Y \mid do(\pi)]$. 

Published as a conference paper at ICLR 2023
For example, for the frontdoor diagram $\mathcal{G}$ in Fig. 2a, the manipulated diagram $\mathcal{G}_S$ with scope $S = \{(X, \emptyset)\}$ is described in Fig. 2b. Since $Z_{Y} = \text{An}(Y)_{\mathcal{G}_S} = \{X, Z, Y\}$, $C_{Y}$ is thus given by $\{X, Y\}$. Lem. 1 implies that $X^* = Pa\{(X, Y)\} \cap \{X\}$ and $Z^* = Pa\{(X, Y)\} \setminus \{X, Y\} = \{Z\}$. Applying IDENTIFY($\mathcal{G}, \emptyset; \{(X, \emptyset)\}$) returns the frontdoor adjustment formula in Eq. (8).

### 3.1 Searching for Identifiable Policy Scopes

The remainder of this section describes an effective algorithm to find identifiable policy scopes $S$ had the latent reward signal $Y$ been observed. Let $\mathbb{S}$ denote the collection of all identifiable policy scopes $S$ from distribution $P(O, Y)$ in the causal diagram $\mathcal{G}$. Our algorithm LISTIDSCOPE, described in Alg. 1, enumerates elements in $\mathbb{S}$. It takes as input a causal diagram $\mathcal{G}$, a reward signal $Y$, and subsets $L = \emptyset$ and $R = \bigcup_{i=1}^{n} PA_i^*$. More specifically, LISTIDSCOPE maintains two scopes $S_l \subseteq S_r$ (Step 2). It performs backtracking search to find identifiable scopes $S$ in $\mathcal{G}$ such that $S_l \subseteq S \subseteq S_r$. It aborts branches that either (1) all subscopes in $S_r$ are identifiable (Step 3); or (2) all subscopes containing $S_l$ are non-identifiable (Step 6). The following proposition supports our aborting criterion.

**Lemma 2.** Given a causal diagram $\mathcal{G}$, for policy scopes $S^r \subseteq S$, $S^r$ is identifiable from distribution $P(O, Y)$ in $\mathcal{G}$ if $S$ is identifiable from $P(O, Y)$ in $\mathcal{G}$.

At Step 7, LISTIDSCOPE picks an arbitrary variable $V$ that is included in input covariates $R$ but not in $L$. It then recursively returns all identifiable policy scopes $S$ in $\mathcal{G}$: the first recursive call returns scopes taking $V$ as an input for some actions $X_i \in X$ and the second call return all scopes that do not consider $V$ when selecting values for all actions $X$. We say a policy $\pi$ is associated with a collection of policy scopes $S$, denoted by $\pi \sim S$, if there exists $S \in \mathbb{S}$ so that $\pi \sim S$. It is possible to show that LISTIDSCOPE produces a collection of identifiable scopes that is sufficient for the imitation task.

**Theorem 3.** For a causal diagram $\mathcal{G}$ and a reward $Y$, LISTIDSCOPE($\mathcal{G}, Y, \emptyset, \bigcup_{i=1}^{n} PA_i^*$) enumerates a subset $\mathbb{S}^r \subseteq \mathbb{S}$ so that for any $\pi \sim \mathbb{S}$, there is $\pi^* \sim \mathbb{S}^r$ where $\mathbb{E}(Y \mid do(\pi^*)) = \mathbb{E}(Y \mid do(\pi^*))$.

Moreover, LISTIDSCOPE outputs identifiable policy scopes with a polynomial delay. This follows from the observation that LISTIDSCOPE searches over a tree of policy scopes with height at most $\bigcup_{i=1}^{n} PA_i^*$ and IDENTIFY($\mathcal{G}, Y, S$) terminates in polynomial steps w.r.t. the size of diagram $\mathcal{G}$.

### 4 Experiments

In this section, we demonstrate our framework on various imitation learning tasks, ranging from synthetic causal models to real-world datasets, including highway driving (Krajewski et al., 2018) and images (LeCun, 1998). We find that our approach is able to incorporate parametric knowledge about the reward function and achieve effective imitating policies across different causal diagrams. For all experiments, we evaluate our proposed Causal-IRL based on the canonical estimation formulation in Eq. (5). As a baseline, we also include: (1) standard BC mimicking the expert’s nominal behavior policy; (2) standard IRL utilizing all observed covariates preceding every $X_i \in X$ while being blind to causal relationships in the underlying model; and (3) Causal-BC (Zhang et al., 2020; Kumor et al., 2021) that learn an imitating policy with the sequential $\pi$-backdoor criterion. We refer readers to (Ruan et al., 2023 Appendix D) for additional experiments and more discussions on the experimental setup.

**Backdoor** Consider an SCM instance compatible with Fig. 1c including binary observed variables $Z_1, X_1, Z_2, X_2, Y \in \{0, 1\}$. Causal-BC utilizes a sequential $\pi$-backdoor admissible scope $\{(X_1, \emptyset), (X_2, \{Z_2\})\}$; while Causal-IRL utilizes the scope $\{(X_1, \emptyset), (X_2, \{Z_2\})\}$ satisfying the minimal sequential $\pi$-backdoor. Simulation results, shown in Fig. 3a, reveal that Causal-IRL consistently outperforms the expert’s policy and other imitation strategies by exploiting additional parametric knowledge about the expected reward $\mathbb{E}(Y \mid X_1, X_2, Z_2)$; Causal-BC is able to achieve the expert’s performance. Unsurprisingly, neither BC nor IRL is able to obtain an imitating policy.
Highway Driving We consider a learning scenario where the agent learns a driving policy from the observed trajectories of a human expert. Causal diagram of this example is provided in (Ruan et al., 2023, Appendix D, Fig. 4) where $X_1$ is the accelerations of the ego vehicle at the previous step; $Z_1$ is both longitudinal and lateral historical accelerations of the ego vehicle two steps ago; $X_2$ is the velocity of the ego vehicle; $Z_2$ is the velocity of the preceding vehicle; $W$ indicates the information from surrounding vehicles. Values of $X_1, X_2, Z_1, Z_2$ are drawn from a real-world driving dataset HighD [Krajewski et al. (2018)]. The reward $Y$ is decided by a non-linear function $f_Y(X_2, Z_2, U_Y)$. Both Causal-IRL and Causal-BC utilize the scope $\{\langle X_1, \emptyset \rangle, \langle X_2, \{Z_2\} \rangle\}$. Causal-IRL also exploits the additional knowledge that the expected reward $\mathbb{E}[Y | X_1, X_2, Z_2]$ is a monotone function via reward augmentation (Li et al., 2017). Simulation results are shown in Fig. 3b. We found that Causal-IRL performs the best among all strategies. Causal-BC is able to achieve the expert’s performance. BC and IRL perform the worst among all and fail to obtain an imitating policy.

MNIST Digits Consider again the frontdoor diagram in Fig. 2a. To evaluate the performance of our proposed approach in high-dimensional domains, we now replace variable $Z$ with sampled images drawn from MNIST digits dataset [LeCun, 1998]. The reward $Y$ is decided by a linear function taking $Z$ and an unobserved confounder $U_{X,Y}$ as input. The Causal-IRL formulates the imitation problem as a two-person zero-sum game through the frontdoor adjustment described in Eq. (9), which can be solved by the MW algorithm (Freund & Schapire, 1999; Syed & Schapire, 2008). As shown in Fig. 3c, simulation results reveal that Causal-IRL outperforms Causal-BC and BC; while IRL performs the worst among all the algorithms.

Infinite MDPUC To demonstrate our proposed framework in the sequential decision-making setting with an infinite horizon, we consider a generalized Markov decision process incorporating unobserved confounders (Ruan & Di, 2022), called the MDPUC (Zhang & Bareinboim, 2022). This sequential model simulates real-world driving dynamics. By exploiting the Markov property over time steps, we are able to decompose the causal diagram over the infinite horizon into a collection of sub-graphs, one for each time step $i = 1, 2, \ldots$. Fig. 1d shows the causal diagram spanning time steps $i = 1, 2, 3$. As a comparison, BC and IRL still utilize the stationary policy $\{\{X_i, \{Z_i\}\}\}$. By applying Thm. 1 at each time step, we obtain a $\pi$-backdoor admissible policy scope $\{\{X_i, \{Z_i, X_{i-1}, Z_{i-1}\}\}\}$ for Causal-IRL and Causal-BC. Simulation results are shown in Fig. 3d. One could see by inspection that Causal-IRL performs the best and achieves the expert’s performance.

5 Conclusion

This paper investigates imitation learning via inverse reinforcement learning (IRL) in the semantical framework of structural causal models. The goal is to find an effective imitating policy that performs at least as well as the expert’s behavior policy from combinations of demonstration data, qualitative knowledge the data-generating mechanisms represented as a causal diagram, and quantitative knowledge about the reward function. We provide a graphical criterion (Thm. 1) based on the sequential backdoor, which allows one to obtain an imitating policy by solving a canonical optimization equation of causal IRL. Such a canonical formulation addresses the challenge of the presence of unobserved confounders (UCs), and is solvable by leveraging standard IRL algorithms (Props. 1 and 2). Finally, we move beyond the backdoor criterion and show that the canonical equation is achievable whenever expected rewards of policies are identifiable had the reward also been observed (Thms. 2 and 3).
ACKNOWLEDGEMENTS

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

ETHICS STATEMENT

This paper investigates the theoretical framework of causal inverse RL from the natural trajectories of an expert demonstrator, even when the reward signal is unobserved. Input covariates used by the expert to determine the original values of the action are unknown, introducing unobserved confounding bias in demonstration data. Our framework may apply to various fields in reality, including autonomous vehicle development, industrial automation, and chronic disease management. A positive impact of this work is that we discuss the potential risk of training IRL policy from demonstrations with the presence of unobserved confounding (UC). Our formulation of causal IRL is inherently robust against confounding bias. For example, solving the causal IRL problem in Eq. (1) requires the imitator to learn an effective policy that maximizes the reward in a worst-case causal model where the performance gap between the expert and imitator is the largest possible. More broadly, automated decision systems using causal inference methods prioritize safety and robustness during their decision-making processes. Such requirements are increasingly essential since black-box AI systems are prevalent, and our understandings of their potential implications are still limited.

REPRODUCIBILITY STATEMENT

The complete proof of all theoretical results presented in this paper, including Thms. 1 and 2 is provided in (Ruan et al., 2023, Appendix B). Details on the implementation of the proposed algorithms are included (Ruan et al., 2023, Appendix C). Finally, (Ruan et al., 2023, Appendix D) provides a detailed description of the experimental setup. Readers could find all appendices as part of the supplementary text after “References” section. We provided references to all existing datasets used in experiments, including HIGHD (Krajewski et al., 2018) and MNIST (LeCun, 1998). Other experiments are synthetic and do not introduce any new assets. Source codes for all experiments and simulations are released in the complete technical report (Ruan et al., 2023).

REFERENCES


