Causal Imitation Learning with Unobserved Confounders

Junzhe Zhang
Columbia University
junzhez@cs.columbia.edu

Daniel Kumor
Purdue University
dkumor@purdue.edu

Elias Bareinboim
Columbia University
eb@cs.columbia.edu

Abstract

One of the common ways children learn is by mimicking adults. Imitation learning focuses on learning policies with suitable performance from demonstrations generated by an expert, with an unspecified performance measure, and unobserved reward signal. Popular methods for imitation learning start by either directly mimicking the behavior policy of an expert (behavior cloning) or by learning a reward function that prioritizes observed expert trajectories (inverse reinforcement learning). However, these methods rely on the assumption that covariates used by the expert to determine her/his actions are fully observed. In this paper, we relax this assumption and study imitation learning when sensory inputs of the learner and the expert differ. First, we provide a non-parametric, graphical criterion that is complete (both necessary and sufficient) for determining the feasibility of imitation from the combinations of demonstration data and qualitative assumptions about the underlying environment, represented in the form of a causal model. We then show that when such a criterion does not hold, imitation could still be feasible by exploiting quantitative knowledge of the expert trajectories. Finally, we develop an efficient procedure for learning the imitating policy from experts’ trajectories.

1 Introduction

A unifying theme of Artificial Intelligence is to learn a policy from observations in an unknown environment such that a suitable level of performance is achieved [33, Ch. 1.1]. Operationally, a policy is a decision rule that determines an action based on a certain set of covariates; observations are possibly generated by a human demonstrator following a different behavior policy. The task of evaluating policies from a combination of observational data and assumptions about the underlying environment has been studied in the literature of causal inference [29] and reinforcement learning [37]. Several criteria, algorithms, and estimation methods have been developed to solve this problem [29, 36, 3, 6, 35, 32, 44]. In many applications, it is not clear which performance measure the demonstrator is (possibly subconsciously) optimizing. That is, the reward signal is not labeled and accessible in the observed expert’s trajectories. In such settings, the performance of candidate policies is not uniquely discernible from the observational data due to latent outcomes, even when infinitely many samples are gathered, complicating efforts to learn policy with satisfactory performance.

An alternative approach used to circumvent this issue is to find a policy that mimics a demonstrator’s behavior, which leads to the imitation learning paradigm [2, 9, 14, 28]. The expectation (or rather hope) is that if the demonstrations are generated by an expert with near-optimal reward, the performance of the imitator would also be satisfactory. Current methods of imitation learning can be categorized into behavior cloning [45, 31, 23, 24, 22] and inverse reinforcement learning [25, 1, 38, 46]. The former focuses on learning a nominal expert policy that approximates the conditional distribution observed input covariates of the behavior policy to the action domain. The latter attempts to learn a reward function that prioritizes observed behaviors of the expert; reinforcement learning methods are then applied using the learned reward function to obtain a nominal performance...
This example shows that even when one is able to perfectly mimic an optimal demonstrator, the fully observed \cite{7}, while another assumed that the primary outcome is observed (e.g., \(Y\) in Fig. 18 \cite{8}).

Finally, our results on high-dimensional, synthetic datasets. For the sake of space constraints, we provide all finding an imitating policy through explicit parametrization of the causal model, and use it to validate knowledge in the observational distribution. (3) We provide an efficient and practical procedure for identifying an imitating policy when the given criterion does not hold, by leveraging the quantitative approach would estimate the conditional distribution \(P(x|z)\). Fig. 1a describes the graphical representation of this environment. A naïve analysis reveals that this naive “cloning” approach leads to sub-optimal performance. Consider an instance where variables \(X, Y, Z, L, U\) are independent variables drawn uniformly over \([0, 1]; \oplus\) represents the exclusive-or operator. The expected reward \(\mathbb{E}[Y|\text{do}(\pi)]\) induced by \(\pi(x|z) = P(x|z)\) is equal to 0.5, which is quite far from the optimal demonstrator’s performance, \(\mathbb{E}[Y] = 1\).

This example shows that even when one is able to perfectly mimic an optimal demonstrator, the learned policy can still be suboptimal. In this paper, we try to explicate this phenomenon and, more broadly, understand imitability through a causal lens\footnote{Some recent progress in the field of causal imitation has been reported, albeit oblivious to the phenomenon described above and our contributions. Some work considered settings in which the input to the expert policy is fully observed \cite{7}, while another assumed that the primary outcome is observed (e.g., \(Y\) in Fig. 18 \cite{8}).}. Our task is to learn an imitating policy that achieves the expert’s performance from demonstration data in a structural causal model \cite{29, Ch. 7}, allowing for unobserved confounders (UCs) affecting both action and outcome variables. Specifically, our contributions are summarized as follows. (1) We introduce a complete graphical criterion for determining the feasibility of imitation from demonstration data and qualitative knowledge about the data-generating process represented as a causal graph. (2) We develop a sufficient algorithm for identifying an imitating policy when the given criterion does not hold, by leveraging the quantitative knowledge in the observational distribution. (3) We provide an efficient and practical procedure for finding an imitating policy through explicit parametrization of the causal model, and use it to validate our results on high-dimensional, synthetic datasets. For the sake of space constraints, we provide all proofs in the complete technical report \cite{15, Appendix A}.

1.1 Preliminaries

In this section, we introduce the basic notations and definitions used throughout the paper. We use capital letters to denote random variables (\(X\)) and small letters for their values (\(x\)). \(\mathcal{X}\) represents the domain of \(X\) and \(\mathcal{P}_X\) the space of probability distributions over \(\mathcal{X}\). For a set \(X, |X|\) denotes its dimension. We consistently use the abbreviation \(P(x)\) to represent the probabilities \(P(X = x)\). Finally, \(I(z = z)\) is an indicator function that returns 1 if \(Z = z\) holds true; otherwise 0.
We write \( Pa \) the subgraph obtained from \( G \) by removing arrows coming into nodes in \( X \); \( G_X \) is a subgraph of \( G \) by removing arrows going out of \( X \). We will use standard family conventions for graphical relationships such as parents, children, descendants, and ancestors. For example, the set of parents of \( X \) in \( G \) is denoted by \( Pa(X)_G = \cup_{x \in X} pa(X)_G \). ch, de and an are similarly defined.

We write \( Pa, Ch, De, An \) if arguments are included as well, e.g. \( De(X)_G = de(X)_G \cup X \). A path from a node \( X \) to a node \( Y \) in \( G \) is a sequence of edges which does not include a particular node more than once. Two sets of nodes \( X, Y \) are said to be d-separated by a third set \( Z \) in a DAG \( G \), denoted by \((X \perp \perp Y | Z)_G\), if every edge path from nodes in one set to nodes in another are “blocked”. The criterion of blockage follows \([29 \text{ Def. 1.2.3}]\).

The basic semantic framework of our analysis rests on variables all possible policies \( M \pi \) of \( X \) independently following the expected value of a reward variable \( Y \) to decide the value of an action variable \( X \). This indicates the presence of an unobserved confounder (UC) affecting both \( f \) and \( L \) represent latent variables, diagram \( G \) a subgraph of \( G \) is a sequence of edges which does not include a particular node more than once. Two sets of nodes \( X, Y \) are said to be d-separated by a third set \( Z \) in a DAG \( G \), denoted by \((X \perp \perp Y | Z)_G\), if every edge path from nodes in one set to nodes in another are “blocked”. The criterion of blockage follows \([29 \text{ Def. 1.2.3}]\).

### 2 Imitation Learning in Structural Causal Models

In this section, we formalize and study the imitation learning problem in causal language. We first define a special type of SCM that explicitly allows one to model the unobserved nature of some endogenous variables, which is called the partially observable structural causal model (POSCM)

**Definition 1 (Partially Observable SCM).** A POSCM is a tuple \( \langle M, O, L \rangle \), where \( M \) is a SCM \( \langle U, V, F, P(u) \rangle \) and \( O \) is a pair of subsets forming a partition over \( V \) (i.e., \( V = O \cup L \) and \( O \cap L = \emptyset \)); \( O \) and \( L \) are called observed and latent endogenous variables, respectively.

Each POSCM \( M \) induces a probability distribution over \( V \), of which one can measure the observed variables \( O \). \( P(o) \) is usually called the observational distribution. \( M \) is associated with a causal diagram \( G \) (e.g., see Fig. 1) where solid nodes represent observed variables \( O \), dashed nodes represent latent variables \( L \), and arrows represent the arguments \( P_{AV} \) of each functional relationship \( f_V \). Exogenous variables \( U \) are not explicitly shown; a bi-directed arrow between nodes \( V_i \) and \( V_j \) indicates the presence of an unobserved confounder (UC) affecting both \( V_i \) and \( V_j \), i.e., \( U_{ij} \cap U_{ji} \neq \emptyset \).

Consider a POSCM \( \langle M, O, L \rangle \) with \( M = \langle U, V, F, P(u) \rangle \). Our goal is to learn an efficient policy \( \pi \) to decide the value of an action variable \( X \in O \). The performance of the policy is evaluated using the expected value of a reward variable \( Y \). Throughout this paper, we assume that reward \( Y \) is latent and \( X \) affects \( Y \) (i.e., \( Y \in L \cap De(X)_G \)). A policy \( \pi \) is a function mapping from values of covariates \( Pa^{*} \subseteq O \setminus De(X)_G \) to a probability distribution over \( X \), which we denote by \( \pi(x|Pa^{*}) \). An intervention following a policy \( \pi \), denoted by \( do(\pi) \), is an operation that draws values of \( X \) independently following \( \pi \), regardless of its original (natural) function \( f_X \). Let \( M_\pi \) denote the manipulated SCM of \( M \) induced by \( do(\pi) \). Similar to atomic settings, the interventional distribution \( P(v|do(\pi)) \) is defined as the distribution over \( V \) in the manipulated model \( M_\pi \), given by:

\[
P(v|do(\pi)) = \sum_u P(u) \prod_{V \in V \setminus \{X\}} P(v|Pa_V, u_V)\pi(x|Pa^*). \tag{1}
\]

The expected reward of a policy \( \pi \) is thus given by the causal effect \( \mathbb{E}[Y|do(\pi)] \). The collection of all possible policies \( \pi \) defines a policy space, denoted by \( \Pi = \{ \pi : \mathcal{P}_{Pa^*} \rightarrow \mathcal{P}_X \} \) (if \( Pa^* = \emptyset \), \( \Pi = \{ \pi : \mathcal{P}_X \} \)). For convenience, we define function \( Pa(\Pi) = Pa^* \). A policy space \( \Pi' \) is a

---

\( G_X \) is a causal diagram associated with the submodel \( M_x \) induced by intervention \( do(x) \).
subspace of \( \Pi \) if \( Pa(\Pi') \subseteq Pa(\Pi) \). We will consistently highlight action \( X \) in dark red, reward \( Y \) in dark blue and covariates \( Pa(\Pi) \) in light red. For instance, in Fig. 14 the policy space over action \( X \) is given by \( \Pi = \{ \pi : \mathcal{G}_Z \mapsto \mathcal{X} \}; \). \( Y \) represents the reward; \( \Pi' = \{ \pi : \mathcal{X} \} \) is a subspace of \( \Pi \).

Our goal is to learn an efficient policy \( \pi \in \Pi \) that achieves satisfactory performance, e.g., larger than a certain threshold \( \mathbb{E}[Y|do(\pi)] \geq \tau \), without knowledge of underlying system dynamics, i.e., the actual, true POSCM \( M \). A possible approach is to identify the expected reward \( \mathbb{E}[Y|do(\pi)] \) for each policy \( \pi \in \Pi \) from the combinations of the observed data \( P(o) \) and the causal diagram \( \mathcal{G} \).

Optimization procedures are applicable to find a satisfactory policy \( \pi \). Let \( \mathcal{M}(\mathcal{G}) \) denote a hypothesis class of POSCMs that are compatible with a causal diagram \( \mathcal{G} \). We define the non-parametric notion of identifiability in the context of POSCMs and conditional policies, adapted from [29, Def. 3.2.4].

**Definition 2** (Identifiability). Given a causal diagram \( \mathcal{G} \) and a policy space \( \Pi \), let \( Y \) be an arbitrary subset of \( \mathcal{V} \). \( P(y|do(\pi)) \) is said to be identifiable w.r.t. \( \langle \mathcal{G}, \Pi \rangle \) if \( P(y|do(\pi); M) \) is uniquely computable from \( P(o; M) \) and \( \pi \) for any POSCM \( M \in \mathcal{M}(\mathcal{G}) \) and any \( \pi \in \Pi \).

In imitation learning settings, however, reward \( Y \) is often not specified and remains latent, which precludes approaches that attempt to identify \( \mathbb{E}[Y|do(\pi)] \):

**Corollary 1.** Given a causal diagram \( \mathcal{G} \) and a policy space \( \Pi \), let \( Y \) be an arbitrary subset of \( \mathcal{V} \). If not all variables in \( Y \) are observed (i.e., \( Y \cap \mathcal{L} \neq \emptyset \)), \( P(y|do(\pi)) \) is not identifiable.

In other words, Corol. 1 shows that when the reward \( Y \) is latent, it is infeasible to uniquely determine values of \( \mathbb{E}[Y|do(\pi)] \) from \( P(o) \). A similar observation has been noted in [20, Prop. 1]. This suggests that we need to explore learning through other modalities.

### 2.1 Causal Imitation Learning

To circumvent issues of non-identifiability, a common solution is to assume that the observed trajectories are generated by an “expert” demonstrator with satisfactory performance \( \mathbb{E}[Y] \), e.g., no less than a certain threshold \( \mathbb{E}[Y] \geq \tau \). If we could find a policy \( \pi \) that perfectly “imitates” the expert with respect to reward \( Y \), \( \mathbb{E}[Y|do(\pi)] = \mathbb{E}[Y] \), the performance of the learner is also guaranteed to be satisfactory. Formally,

**Definition 3** (Imitability). Given a causal diagram \( \mathcal{G} \) and a policy space \( \Pi \), let \( Y \) be an arbitrary subset of \( \mathcal{V} \). \( P(y|do(\pi)) \) is said to be imitable w.r.t. \( \langle \mathcal{G}, \Pi \rangle \) if there exists a policy \( \pi \in \Pi \) uniquely computable from \( P(o; M) \) such that \( P(y|do(\pi); M) = P(y; M) \) for any POSCM \( M \in \mathcal{M}(\mathcal{G}) \).

Our task is to determine the imitability of the expert performance. More specifically, we want to learn an imitating policy \( \pi \in \Pi \) from \( P(o) \) such that \( P(y|do(\pi)) = P(y; M) \) in any POSCM \( M \) associated with the causal diagram \( \mathcal{G} \). Consider Fig. 3a as an example. \( P(y) \) is imitable with policy \( \pi(x) = P(x) \) since by Eq. (1) and marginalization, \( P(y|do(\pi)) = \sum_{x,w} P(y|w)P(w|x)\pi(x) = \sum_{x,w} P(y|w)P(w)P(x) = P(y) \). In practice, unfortunately, the expert’s performance cannot always be imitated. To understand this setting, we first write, more explicitly, the conditions under which this is not the case:

**Lemma 1.** Given a causal diagram \( \mathcal{G} \) and a policy space \( \Pi \), let \( Y \) be an arbitrary subset of \( \mathcal{V} \). \( P(y|do(\pi)) \) is not imitable w.r.t. \( \langle \mathcal{G}, \Pi \rangle \) if there exists two POSCMs \( M_1, M_2 \in \mathcal{M}(\mathcal{G}) \) satisfying \( P(o; M_1) = P(o; M_2) \) while there exists no policy \( \pi \in \Pi \) such that for \( i = 1, 2 \), \( P(y|do(\pi); M_i) = P(y; M_i) \).

It follows as a corollary that \( P(y) \) is not imitable if there exists a POSCM \( M \) compatible with \( \mathcal{G} \) such that no policy \( \pi \in \Pi \) could ensure \( P(y|do(\pi); M) = P(y; M) \). For instance, consider the causal diagram \( \mathcal{G} \) and policy space \( \Pi \) in Fig. 3b. Here, the expert’s reward \( P(y) \) is not imitable: consider a POSCM with functions \( X \leftarrow U, W \leftarrow X, Y \leftarrow U \oplus \neg W \); values \( U \) are drawn uniformly over \( \{0, 1\} \). In this model, \( P(Y = 1|do(\pi)) = 0.5 \) for any policy \( \pi \), which is far from the optimal expert reward, \( P(Y = 1) = 1 \).

An interesting observation from the above example of Fig. 3b is that the effect \( P(y|do(\pi)) \) is identifiable, following the front-door criterion in [29, Thm. 3.3.4], but no
The imitability of \( P(y) \) has been previously shown; while \( P(y|\text{do}(\pi)) \) is not identifiable due to latent reward \( Y \) (Corol. \[1\]).

In general, the problem of imitability is orthogonal to identifiability, and, therefore, requires separate consideration. Since imitability does not always hold, we introduce a useful graphical criterion for determining whether imitating an expert’s performance is feasible, and if so, how.

**Theorem 1 (Imitation by Direct Parents).** Given a causal diagram \( \mathcal{G} \) and a policy space \( \Pi \), \( P(y) \) is imitable w.r.t. \( (\mathcal{G}, \Pi) \) if \( \text{pa}(X)_\mathcal{G} \subseteq \text{pa}(\Pi) \) and there is no bi-directed arrow pointing to \( X \) in \( \mathcal{G} \). Moreover, the imitating policy \( \pi \in \Pi \) is given by \( \pi(x|\text{pa}(\Pi)) = P(x|\text{pa}(X)_\mathcal{G}) \).

In words, Thm. \[1\] says that if the expert and learner share the same policy space, then the policy is always imitable. In fact, this result can be seen as a causal justification for when the method of “behavior cloning”, widely used in practice, is valid, leading to proper imitation. When the original behavior policy \( f_X \) is contained in the policy space \( \Pi \), the learner could imitate the expert’s reward \( P(y) \) by learning a policy \( \pi \in \Pi \) that matches the distribution \( P(x|\text{pa}(\Pi)) \) \[45, 31\].

**Corollary 2.** Given a causal diagram \( \mathcal{G} \), a policy space \( \Pi \), and an observational distribution \( P(o) \), let \( Y \) be an arbitrary subset of \( V \). \( P(y) \) is said to be practically imitable (for short, p-imitable) w.r.t. \( (\mathcal{G}, \Pi, P(o)) \) if there exists a policy \( \pi \in \Pi \) uniquely computable from \( P(o) \) such that \( P(y|\text{do}(\pi); M) = P(y; M) \) for any POSCM \( M \in \mathcal{M}_{\mathcal{G}, P} \).

The following corollary can be derived based on the definition of practical imitability.

**Corollary 2.** Given a causal diagram \( \mathcal{G} \), a policy space \( \Pi \) and an observational distribution \( P(o) \), let a subset \( Y \subseteq V \). If \( P(y) \) is imitable w.r.t. \( (\mathcal{G}, \Pi, P(o)) \), \( P(y) \) is p-imitable w.r.t. \( (\mathcal{G}, \Pi, P(o)) \).

Compared to Def. \[3\] the practical imitability of Def. \[5\] aims to find an imitating policy for a subset of candidate POSCMs \( \mathcal{M}_{\mathcal{G}, P} \) restricted to match a specific observational distribution \( P(o) \). Def. \[3\] on
the other hand, requires only the causal diagram $G$. In other words, for an expert’s performance $P(y)$ that is non-imitable w.r.t. $(G, \Pi)$, it could still be p-imitable after analyzing actual probabilities of the observational distribution $P(o)$.

For concreteness, consider again $P(y)$ in Fig. 3b which is not imitable due to the bi-directed arrow $X \leftrightarrow Y$. However, new imitation opportunities arise when actual parameters of the observational distribution $P(x, w, y)$ are provided. Suppose the underlying POSCM is given by: $X \leftarrow U_X \odot U_Y$, $W \leftarrow X \odot W, Y \leftarrow W \odot U_Y$ where $U_X, U_Y, U_W$ are independent binary variables drawn from $P(U_X = 1) = P(U_Y = 1) = P(U_W = 0) = 0.9$. Here, the causal effect $P(y|do(x))$ is identifiable from $P(x, w, y)$ following the front-door formula $P(y|do(x)) = \sum_{w} P(w|x) \sum_{x'} P(y|w, x')$. Consider as an example Fig. 1c where the supergraph $G \cup \Pi$ is a supergraph of $P$. For example, in Fig 1d, $S$ is a supergraph of $P$. For each policy $\pi$, $P(\pi)$ is an $\pi$-imitable with atomic intervention $do(X = 0)$. In the most practical settings, the expert reward $P(y)$ rarely equates to $P(y|do(x))$; stochastic policies $\pi(x)$ are then applicable to imitate $P(y)$ by re-weighting $P(y|do(x))$ induced by the corresponding atomic interventions $\pi$ to tackle p-imitability in a general way, we proceed by defining a set of observed variables that serve as a surrogate of the unobserved $Y$ with respect to interventions on $X$. Formally,

**Definition 6 (Imitation Surrogate).** Given a causal diagram $G$, a policy space $\Pi$, let $S$ be an arbitrary subset of $O$. $S$ is an imitation surrogate (for short, surrogate) w.r.t. $(G, \Pi)$ if $(Y \perp \!\!\!\perp X|S)_{G\cup\Pi}$ where $G \cup \Pi$ is a supergraph of $G$ by adding arrows from $Pa(\Pi)$ to $X$; $X$ is a new parent to $X$.

An surrogate $S$ is said to be minimal if there exists no subset $S' \subset S$ such that $S'$ is also a surrogate w.r.t. $(G, \Pi)$. Consider as an example Fig. 1c where the supergraph $G \cup \Pi$ coincides with the causal diagram $G$. By Def. 6 both $\{W, S\}$ and $\{S\}$ are valid surrogate relative to $(X, Y)$ with $\{S\}$ being the minimal one. By conditioning on $S$, the decomposition of Eq. 1 implies $P(y|do(\pi)) = \sum_{x, w, u} P(y|s) P(s|w) P(w|x) P(u) = \sum_{x} P(y|s) P(s|do(\pi))$. That is, the surrogate $S$ mediates all influence of interventions on action $X$ to reward $Y$. It is thus sufficient to find an imitating policy $\pi$ such that $P(s|do(\pi)) = P(s)$ for any POSCM $M$ associated with Fig. 1c.

The result policy is guaranteed to imitate the expert’s reward $P(y)$.

When a surrogate $S$ is found and $P(s|do(\pi))$ is identifiable, one could compute $P(s|do(\pi))$ for each policy $\pi$ and check if it matches $P(s)$. In many settings, however, $P(s|do(\pi))$ is not identifiable w.r.t. $(G, \Pi)$. For example, in Fig 1d, $S$ is a surrogate w.r.t. $(G, \Pi)$, but $P(s|do(\pi))$ is not identifiable due to collider $Z$ ($\tau$ uses non-descendants as input by default). Fortunately, identifying $P(s|do(\pi))$ may still be feasible in some subspaces of $\Pi$:

**Definition 7 (Identifiable Subspace).** Given a causal diagram $G$, a policy space $\Pi$, and a subset $S \subseteq O$, let $\Pi'$ be a policy subspace of $\Pi$. $\Pi'$ is said to be an identifiable subspace (for short, id-subspace) w.r.t. $(G, \Pi, S)$ if $P(s|do(\pi))$ is identifiable w.r.t. $(G, \Pi')$.

Consider a policy subspace $\Pi' = \{\pi \in \mathcal{P}_X\}$ described in Fig. 1d (i.e. $\pi$ that does not exploit information from covariates $Z$). $P(s|do(\pi))$ is identifiable w.r.t. $(G, \Pi')$ following the front-door adjustment on $W$ [29 Thm. 3.3.4]. We could then evaluate interventional probabilities $P(s|do(\pi))$ for each policy $\pi \in \Pi'$ from the observational distribution $P(x, w, s, z)$; the imitating policy is obtainable by solving the equation $P(s|do(\pi)) = P(s)$. In other words, $(S, \Pi')$ forms an instrument that allows one to solve the imitation learning problem in Fig. 1d.

**Definition 8 (Imitation Instrument).** Given a causal diagram $G$ and a policy space $\Pi$, let $S$ be a subset of $O$ and $\Pi'$ be a subspace of $\Pi$. $(S, \Pi')$ is said to be an imitation instrument (for short, instrument) if $S$ is a surrogate w.r.t. $(G, \Pi')$ and $\Pi'$ is an id-subspace w.r.t. $(G, \Pi, S)$.

**Lemma 2.** Given a causal diagram $G$, a policy space $\Pi$, and an observational distribution $P(o)$, let $(S, \Pi')$ be an instrument w.r.t. $(G, \Pi)$. If $P(s)$ is p-imitable w.r.t. $(G, \Pi', P(o))$, then $P(y)$ is p-imitable w.r.t. $(G, \Pi, P(o))$. Moreover, an imitating policy $\pi$ for $P(s)$ w.r.t. $(G, \Pi', P(o))$ is also imitating policy for $P(y)$ w.r.t. $(G, \Pi, P(o))$.

In words, Lem. 2 shows that when an imitation instrument $(S, \Pi')$ is present, we could reduce the original imitation learning on a latent reward $Y$ to a p-imitability problem over observed surrogate variables $S$ using policies in an identifiable subspace $\Pi'$. The imitating policy $\pi$ is obtainable by solving the equation $P(s|do(\pi)) = P(s)$.

---

3Consider a variation of the model where $P(U_W = 1) = 0.7$. $P(y)$ is p-imitable with $P(X = 0) = 0.75.$
3.1 Confounding Robust Imitation

Our task in this section is to introduce a general algorithm that finds instruments, and learns a p-imitating policy given \(G, \Pi, P(o)\). A naive approach is to enumerate all pairs of subset \(S\) and subspace \(\Pi'\) and check whether they form an instrument; if so, we can compute an imitating policy for \(P(s)\) w.r.t. \(G, \Pi', P(o)\). However, the challenge is that the number of all possible subspaces \(\Pi'\) (or subsets \(S\)) can be exponentially large. Fortunately, we can greatly restrict this search space. Let \(G \cup \{Y\}\) denote a causal diagram obtained from \(G\) by making reward \(Y\) observed. The following proposition suggests that it suffices to consider only identifiable subspaces w.r.t. \(G \cup \{Y\}, \Pi, Y\).

**Lemma 3.** Given a causal diagram \(G\), a policy space \(\Pi\), let a subspace \(\Pi' \subset \Pi\). If there exists \(S \subseteq O\) such that \((S, \Pi')\) is an instrument w.r.t. \((G, \Pi), \Pi'\) is an id-subspace w.r.t. \((G \cup \{Y\}, \Pi, Y)\).

Our algorithm \textsc{Imitate} is described in Alg. 1. We assume access to an \textsc{Identify} oracle \cite{DBLP:journals/tit/Cheng19, DBLP:journals/tit/Cheng20} that takes as input a causal diagram \(G\), a policy space \(\Pi\) and a set of observed variables \(S\). If \(P(s|do(\pi))\) is identifiable w.r.t. \((G, \Pi)\), \textsc{Identify} returns “YES”; otherwise, it returns “NO”. For details about the \textsc{Identify} oracle, we refer readers to \cite{Cheng19} Appendix B]. More specifically, \textsc{Imitate} takes as input a causal diagram \(G\), a policy space \(\Pi\) and an observational distribution \(P(o)\). At Step 2, \textsc{Imitate} applies a subroutine \textsc{ListIdentSPACE} to list identifiable subspaces \(\Pi'\) w.r.t. \((G \cup \{Y\}, \Pi, Y)\), following the observation made in Lem. 3. The implementation details of \textsc{ListIdentSPACE} are provided in \cite{Cheng19} Appendix C]. When an identifiable subspace \(\Pi'\) is found, \textsc{Imitate} tries to obtain a surrogate \(S\) w.r.t. the diagram \(G\) and subspace \(\Pi'\). While there could exist multiple such surrogates, the following proposition shows that it is sufficient to consider only minimal ones.

**Lemma 4.** Given a causal diagram \(G\), a policy space \(\Pi\), an observational distribution \(P(o)\) and a subset \(S \subseteq O\). \(P(s)\) is p-imitable only if for any \(S' \subseteq S\), \(P(s')\) is p-imitable w.r.t. \((G, \Pi, P(o))\).

We apply a subroutine \textsc{ListMinSep} in \cite{DBLP:journals/tit/Cheng20} to enumerate minimal surrogates in \(O\) that d-separate \(X\) and \(Y\) in the supergraph \(G \cup \Pi'\). When a minimal surrogate \(S\) is found, \textsc{Imitate} uses the \textsc{Identify} oracle to validate if \(P(s|do(\pi))\) is identifiable w.r.t. \((G, \Pi')\), i.e., \((S, \Pi')\) form an instrument. Consider Fig. 1d as an example. While \(P(y|do(\pi))\) is not identifiable for every policy in \(\Pi\) had \(Y\) been observed, \(\Pi\) contains an id-subspace \(\{\pi: \mathcal{X}\}\) w.r.t. \((G \cup \{Y\}, \Pi, Y)\), which is associated with a minimal surrogate \(S\). Applying \textsc{Identify} confirms that \((S, \{\pi: \mathcal{X}\})\) is an instrument.

At Step 5, \textsc{Imitate} solves for a policy \(\pi\) in the subspace \(\Pi'\) that imitates \(P(s)\) for all instances in the hypothesis class \(\mathcal{H}(G, P)\). If such a policy exists, \textsc{Imitate} returns \(\pi\); otherwise, the algorithm continues. Since \((S, \Pi')\) is an instrument, Lem. 2 implies that the learned policy \(\pi\), if it exists, is ensured to imitate the expert reward \(P(y)\) for any POSCM \(M \in \mathcal{H}(G, P)\).

**Theorem 3.** Given a causal diagram \(G\), a policy space \(\Pi\), and an observational distribution \(P(o)\), if \textsc{Imitate} returns a policy \(\pi \in \Pi, P(y)\) is p-imitable w.r.t. \((G, \Pi, P(o))\). Moreover, \(\pi\) is an imitating policy for \(P(y)\) w.r.t. \((G, \Pi, P(o))\).

3.2 Optimizing Imitating Policies

We now introduce optimization procedures to solve for an imitating policy at Step 5 of \textsc{Imitate} algorithm. Since the pair \((S, \Pi')\) forms a valid instrument (ensured by Step 4), the interventional distribution \(P(s|do(\pi); M)\) remains invariant among all models in \(\mathcal{H}(G)\), i.e., \(P(s|do(\pi))\) is identifiable w.r.t. \((G, \Pi)\). We could thus express \(P(s|do(\pi); M)\) for any \(M \in \mathcal{H}(G, P)\) as a function of the observational distribution \(P(o)\); for simplicity, we write \(P(s|do(\pi)) = P(s|do(\pi); M)\). The imitating policy \(\pi\) is obtainable by solving the equation \(P(s|do(\pi)) = P(s)\). We could derive a closed-form formula for \(P(s|do(\pi))\) following standard causal identification algorithms in \cite{DBLP:journals/tit/Cheng19, DBLP:journals/tit/Cheng20}. As an
example, consider again the setting of Fig. 1c with binary $X, W, S, Z$; parameters of $P(x, w, s, z)$ could be summarized using an 8-entry probability table. The imitating policy $\pi(x)$ is thus a solution of a series of linear equations $\sum_x \pi(x) P(s|do(x)) = P(s)$ and $\sum_x \pi(x) = 1$, given by:

$$
\pi(x_0) = \frac{P(s_1) - P(s_1|do(x_0))}{P(s_1|do(x_1)) - P(s_1|do(x_0))}, \\
\pi(x_1) = \frac{P(s_1|do(x_1)) - P(s_1)}{P(s_1|do(x_1)) - P(s_1|do(x_0))}.
$$

Among quantities in the above equation, $x_i, s_j$ represent assignments $X = i, S = j$ for $i, j \in \{0, 1\}$. The interventional distribution $P(s|do(x))$ could be identified from $P(x, w, s, z)$ using the front-door adjustment formula $P(s|do(x)) = \sum_w P(w|x) \sum_{x'} P(s|x', w) P(x')$ [29 Thm. 3.3.4].

However, evaluating interventional probabilities $P(s|do(\pi))$ from the observational distribution $P(\alpha)$ could be computational challenging if some variables in $O$ are high-dimensional (e.g., $W$ in Fig. 1d). Properties of the imitation instrument $(\langle S, \Pi \rangle)$ suggest a practical approach to address this issue. Since $P(s|do(\pi))$ is identifiable w.r.t. $(G, \Pi')$, by Def. 2 it remains invariant over the models in the hypothesis class $\mathcal{M}(G)$. This means that we could compute interventional probabilities $P(s|do(\pi); \hat{M})$ in an arbitrary model $\hat{M} \in \mathcal{M}(G)$; such an evaluation will always coincide with the actual, true causal effect $P(s|do(\pi))$ in the underlying model. This observation allows one to obtain an imitating policy through the direct parametrization of POSCMs [21]. Let $\mathcal{N}(G)$ be a parametrized subfamily of POSCMs in $\mathcal{M}(G)$. We could obtain an POSCM $\hat{M} \in \mathcal{N}(G)$ such that its observational distribution $P(\alpha; \hat{M}) = P(\alpha)$; an imitating policy $\pi$ is then computed in the parametrized model $\hat{M}$. Corol. 3 shows that such a policy $\pi$ is an imitating policy for the expert’s reward $P(y)$.

**Corollary 3.** Given a causal diagram $G$, a policy space $\Pi$, and an observational distribution $P(\alpha)$, let $(S, \Pi')$ be an instrument w.r.t. $(G, \Pi)$. If there exists a POSCM $M \in \mathcal{M}(G,p)$ and a policy $\pi \in \Pi'$ such that $P(s|do(\pi); M) = P(s)$, then $P(y)$ is p-imitable w.r.t. $(G, \Pi, P(\alpha))$. Moreover, $\pi$ is an imitating policy for $P(y)$ w.r.t. $(G, \Pi, P(\alpha))$.

In practical experiments, we consider a parametrized family of POSCMs $\mathcal{N}(G)$ where functions associated with each observed variable in $O$ are parametrized by a family of neural networks, similar to [21]. Using the computational framework of Generative Adversarial Networks (GANs) [9, 27], we obtain a model $\hat{M} \in \mathcal{N}(G)$ satisfying the observational constraints $P(\alpha; \hat{M}) = P(\alpha)$. The imitating policy is trained through explicit interventions in the learned model $\hat{M}$; a different GAN is then deployed to optimize the policy $\pi$ so that it imitates the observed trajectories drawn from $P(s)$.

### 4 Experiments

We demonstrate our algorithms on several synthetic datasets, including highD [18] consisting of natural trajectories of human driven vehicles, and MNIST digits. In all experiments, we test our causal imitation method (ci); we apply Thm. 2 when there exists an $\pi$-backdoor admissible set; otherwise, Alg. 1 is used to leverage the observational distribution. As a baseline, we also include naïve behavior cloning (bc) that mimics the observed conditional distribution $P(x|pa(II))$, as well as the actual reward distribution generated by an expert (opt). We found that our algorithms consistently imitate distributions over the expert’s reward in imitable (p-imitable) cases; and p-imitable instances commonly exist. We refer readers to [15, Appendix D] for more experiments, details, and analysis.
We investigate the imitation learning in the semantics of structural causal models. The goal is to find with proper training in causal reasoning. A favorable characteristic of causal inference methods is which finds an imitating policy, by exploiting quantitative knowledge contained in the observational This paper investigates the theoretical framework of learning a policy that imitates the distribution of structural causal models [29, Ch. 7], its adoption could cultivate machine learning practitioners confounding, as shown in Sec. 1 and 4. More broadly, since our method is based on the semantics of structural causal models [29, Ch. 7], its adoption could cultivate machine learning practitioners with proper training in causal reasoning. A favorable characteristic of causal inference methods is

Highway Driving We consider a modified example of the drone recordings of human-driven cars in Sec. [1] where the driver’s braking action $W$ of the left-side car is also observed. Fig. 4a shows the causal diagram of this environment; $Z$ represent the velocity of the front-car; action $X$ represents the velocity of the driving car; $W$ and the reward signal $Y$ are both affected by an unobserved confounder $U$, representing the weather condition. In Fig. 4a, $\{Z, W\}$ is $\pi$-backdoor admissible while $\{Z, W\}$ is not due to active path $X \leftarrow L \rightarrow W \leftrightarrow Y$. We obtain policies for the causal and naive imitators training two separate GANs. Distributions $P(y|do(\pi))$ induced by all algorithms are reported in Fig. 4c. We also measure the L1 distance between $P(y|do(\pi))$ and the expert’s reward $P(y)$. We find that the causal approach $(ci)$, using input set $\{Z\}$, successfully imitates $P(y) (L1 = 0.0018)$. As expected, the naive approach $(bc)$ utilizing all covariates $\{Z, W\}$ is unable to imitate the expert $(L1 = 0.2937)$.

MNIST Digits We consider an instance of Fig. 1c where $X, S, Y$ are binary variables; binary values of $W$ are replaced with corresponding images of MNIST digits (pictures of 1 or 0), determined based on the action $X$. For the causal imitator $(ci)$, we learn a POSCM $\hat{M}$ such that $P(x, w, s; \hat{M}) = P(x, w, s)$. To obtain $\hat{M}$, we train a GAN to imitate the observational distribution $P(x, w, s)$, with a separate generator for each $X, W, S$. We then train a separate discriminator measuring the distance between observed trajectories $P(s)$ and interventional distribution $P(s|do(\pi); \hat{M})$ over the surrogate $\{S\}$. The imitating policy is obtained by minimizing such a distance. Distributions $P(y|do(\pi))$ induced by all algorithms are reported in Fig. 4c. We find that the causal approach $(ci)$ successfully imitates $P(y) (L1 = 0.0634)$. As expected, the naive approach $(bc)$ mimicking distribution $P(x)$ is unable to imitate the expert $(L1 = 0.1900)$.

5 Conclusion

We investigate the imitation learning in the semantics of structural causal models. The goal is to find an imitating policy that mimics the expert behaviors from combinations of demonstration data and qualitative knowledge about the data-generating process represented as a causal diagram. We provide a graphical criterion that is complete (i.e., sufficient and necessary) for determining the feasibility of learning an imitating policy that mimics the expert’s performance. We also study a data-dependent notion of imitability depending on the observational distribution. An efficient algorithm is introduced which finds an imitating policy, by exploiting quantitative knowledge contained in the observational data and the presence of surrogate endpoints. Finally, we propose a practical procedure for estimating such an imitating policy from observed trajectories of the expert’s demonstrations.

Broader Impact

This paper investigates the theoretical framework of learning a policy that imitates the distribution over a primary outcome from natural trajectories of an expert demonstrator, even when the primary outcome itself is unobserved and input covariates used by the expert determining original values of the action are unknown. Since in practice, the actual reward is often unspecified and the learner and the demonstrator rarely observe the environment in the same fashion, our methods are likely to increase the progress of automated decision systems. Such systems may be applicable to various fields, including the development of autonomous vehicle, industrial automation and the management of chronic disease. These applications may have a broad spectrum of societal implications. The adoption of autonomous driving and industrial automation systems could save cost and reduce risks such as occupational injuries; while it could also create unemployment. Treatment recommendation in the clinical decision support system could certainly alleviate the stress on the healthcare workers. However, this also raise questions concerning with the accountability in case of medical malpractice; collection of private personal information could also make the hospital database valuable targets for malicious hackers. Overall, we would encourage research to understand the risks arising from automated decision systems and mitigations for its negative impact.

Recently, there is a growing amount of dataset of natural vehicle trajectories like highD [18] being licensed for commercial use. An immediate positive impact of this work is that we discuss potential risk of training decision-making policy from the observational data due to the presence of unobserved confounding, as shown in Sec. [1] and [5]. More broadly, since our method is based on the semantics of structural causal models [29, Ch. 7], its adoption could cultivate machine learning practitioners with proper training in causal reasoning. A favorable characteristic of causal inference methods is
that they are inherently robust: for example, the definition of imitability Def.\ref{def:imitability} requires the imitating policy to perfectly mimics the expert performance in any model compatible to the causal diagram. Automated decision systems using the causal inference methods prioritize the safety and robustness in decision-making, which is increasingly essential since the use of black-box AI systems is prevalent and our understandings of their potential implications are still limited.

**Acknowledge**

The authors were partially supported by grants from NSF IIS-1704352 and IIS-1750807 (CAREER).

**References**


A Proofs

**Corollary 1.** Given a causal diagram $\mathcal{G}$ and a policy space $\Pi$, let $Y$ be an arbitrary subset of $V$. If not all variables in $Y$ are observed (i.e., $Y \cap L \neq \emptyset$), $P(y|do(\pi))$ is not identifiable.

**Proof.** Let $Y \subseteq Y \cap L$. For any SCM $M_1$ that induces $\mathcal{G}$, we could obtain an SCM $M_2$ by replacing $f_Y$ and $P(u_Y)$ associated with $Y$. Since there is no restriction on the parametrization of $f_Y$ and $P(u_Y)$, we could always ensure $P(y|do(\pi); M_1) \neq P(y|do(\pi); M_2)$. For example, in both $M_1, M_2,$ let $Y \leftarrow U_Y$; we define $P(U_Y = 0; M_1) = 0.1$ and $P(U_Y = 0; M_2) = 0.9$. It is immediate to see that $P(y|do(\pi); M_1) \neq P(y|do(\pi); M_2)$, i.e., $P(y|do(\pi))$ is not identifiable. \hfill \Box

**Lemma 1.** Given a causal diagram $\mathcal{G}$ and a policy space $\Pi$, let $Y$ be an arbitrary subset of $V$. $P(y)$ is not imitable w.r.t. $(\mathcal{G}, \Pi)$ if there exist two POSCMs $M_1, M_2 \in \mathcal{M} (\mathcal{G})$ satisfying $P(\alpha; M_1) = P(\alpha; M_2)$ while there exists no policy $\pi \in \Pi$ such that for $i = 1, 2$, $P(y|do(\pi); M_i) = P(y; M_i)$.

**Proof.** The lack of existence of a shared imitating policy between $M_1, M_2$ eliminates the possibility of the existence of a function from $P(\alpha)$ to a policy $\pi$ that imitates $P(y)$ in any POSCM compatible with the causal diagram $\mathcal{G}$. \hfill \Box

**Theorem 1** (Imitation by Direct Parents). Given a causal diagram $\mathcal{G}$ and a policy space $\Pi$, $P(y)$ is imitable w.r.t. $(\mathcal{G}, \Pi)$ if $pa(X)_{G} \subseteq Pa(\Pi)$ and there is no bi-directed arrow pointing to $X$ in $\mathcal{G}$. Moreover, the imitating policy $\pi \in \Pi$ is given by $\pi(x|pa(\Pi)) = P(x|pa(X)_{G})$.

**Proof.** Since there is not bi-directed arrow pointing into $X$, it is variables that $pa(X)_{G}$ is $\pi$-backdoor admissible relative to $(X, Y)$ in $\mathcal{G}$. By Rule 2 of do-calculus, we have

$$P(y) = \sum_{x, pa(X)_{G}} P(pa(X)_{G})P(x|pa(X)_{G})P(y|do(x), pa(X)_{G}).$$

Since $pa(X)_{G} \subseteq Pa(\Pi)$, the conditional distribution $P(x|pa(X)_{G})$ can be represented as a policy in $\Pi$. Let $\pi(x|Pa(\Pi)) = \pi(x|pa(X)_{G}) = P(x|pa(X)_{G})$. We must have

$$P(y) = \sum_{x, pa(X)_{G}} P(pa(X)_{G})\pi(x|pa(X)_{G})P(y|do(x), pa(X)_{G}) = P(y|do(\pi))).$$ \hfill \Box

**Lemma 5.** Given a causal diagram $\mathcal{G}$ and a policy space $\Pi$, let $Z = An(Y)_{G} \cap Pa(\Pi)$. If $P(y)$ is imitable relative to $(\mathcal{G}, \Pi)$, then $(Y \perp \!\!\!\!\perp Z | Z)_{G^* \Pi}$.

**Proof.** We consider first a simplified causal diagram $\mathcal{H}$ where all endogenous variables are observed, i.e., $V = O$; and $Pa(\Pi) = O \setminus \{X, Y\}$. In this diagram $\mathcal{G}$, $(Y \not\perp \!\!\!\!\perp X | Z)_{\mathcal{G}^*}$ implies that there exists a bi-directed path $l$ of the form $X \leftrightarrow Z_1 \leftrightarrow Z_2 \leftrightarrow \cdots \leftrightarrow Z_n \leftrightarrow Y$ such that $Z_i \in Z$ for any $i = 1, \ldots, n$. We denote by $Z^* = \{Z_1, \ldots, Z_n\}$. Recall that $X \in An(Y)_{\mathcal{H}}$. We could thus obtain a subgraph $\mathcal{H}'$ of $\mathcal{H}$ that satisfies the following condition:

- $\mathcal{H}'$ contains the bi-directed path $l$.
- All nodes in $\mathcal{H}'$ are descendants of $Z^* \cup \{X\}$.
- $Y$ is a descendant for all nodes in $\mathcal{H}'$.
- Every endogenous node $V$ in $\mathcal{H}'$ has at most one child.
- All bi-directed arrows in $\mathcal{H}'$ are contained in $l$.

A mental image for depicting $\mathcal{H}'$ is to think of a tree rooted in node $Y$; $X, Z_1, \ldots, Z_n$ are leaf nodes; nodes $X, Z_1, \ldots, Z_n, Y$ are connected by the bi-directed path $l$; $\mathcal{H}'$ contain no bi-directed arrow except path $l$. We now construct a POSCM $M'$ compatible with $\mathcal{G}'$. More specifically, values of each exogenous confounder $U_i$ residing on $l$ are drawn uniformly over a binary domain $\{0, 1\}$. For each endogenous variable $V_i$ in $\mathcal{H}'$, its values is equal to the parity sum of its parents in $\mathcal{H}'$, i.e., $V_i \leftarrow \oplus_{V_j \in pa(V_i)_{M'}} V_j$. By construction, in the subgraph $\mathcal{H}'$, each exogenous $U_i$ has exactly two
directed paths going to $Y$. This means that in the constructed model $M'$, observed values of $Y$ is always equal to 0, i.e., the reward distribution $P(Y = 0; M') = 1$.

Consider a policy space $\Pi'$ of the form $\{\pi : \mathcal{G}_{\pi} \to \mathcal{O}_X\}$. Let $U_n$ denote the exogenous confounder residing on $l$ that is closest to $Y$, i.e., $X \leftrightarrow Z_1 \leftrightarrow Z_2 \leftrightarrow \cdots \leftrightarrow Z_n \leftrightarrow U_n \rightarrow Y$. By the definition of $M'$, values of each variable $Z_i \in Z^*$ is the parity sum of at least two exogenous variables $U, U_n$. Since values of each $U_i$ are drawn uniformly over $\{0, 1\}$, we must have $P(u_i, z^*) = P(u_i)P(z^*)$, or equivalently, $P(u_i | z^*) = P(u_i)$. By definition, we could obtain from $I$ a directed path going from $U_n$ to $Y$ that is not intercepted by $Z^*$. That is, given any value $Z^* = z^*$, values of $Y$ is decided by a parity function taking $U_n$ as an input. Since $P(u_n | z^*) = P(u_n)$ and $U_n$ is drawn uniformly over $\{0, 1\}$, it is verifiable that in $M'$, given any $Z^* = z^*$, the conditional distribution $P(Y = 0 | do(x), z^*; M') = 0.5$. This means that for any policy $\pi \in \Pi'$, the interventional distribution $P(Y = 0 | do(\pi); M') = 0.5$, which is far from the observational distribution $P(Y = 0; M') = 1$. That is, $P(y) = 0$ is not imitable w.r.t. $\langle H', \Pi' \rangle$.

We will now show the non-imitability of $P(y)$ w.r.t. $\langle H, \Pi \rangle$. For any node $V$ that is not included in $\mathcal{H}'$, let its values be decided by an independent noise $U_V$ drawn uniformly over $\{0, 1\}$. We denote this extended POSCM by $M$. Obviously, $M$ is compatible with the causal diagram $\mathcal{H}$. In this model, for any covariate $V \in Pa(\Pi) \setminus Z^*$, it is either (1) a random variable disconnected to any other exogenous variable in $M$; or (2) decided by a function taking $Z^*$ as input. That is, $V \in Pa(\Pi) \setminus Z^*$ contains no value of information with regard to the reward $Y$ when intervening on action $X$. By [17] Ch. 23.6, this means that for any policy $\pi \in \Pi$, there exists a policy $\pi'$ in the subspace $\Pi'$ such that $P(y | do(\pi); M) = P(y | do(\pi'); M)$. Recall that there exists no policy $\pi$ in $\Pi'$ that could ensure $P(y | do(\pi); M) = P(y'; M')$ in POSCM $M'$. By definition of $M$ and $M'$, we must have $P(y; M) = P(y; M')$ and for any policy $\pi \in \Pi$, $P(y | do(\pi); M) = P(y | do(\pi); M')$. It is immediate to see that there exists no policy $\pi \in \Pi'$ that could induce $P(y | do(\pi); M) = P(y; M)$ in the extended POSCM $M$, i.e., $P(y) = 0$ is not imitable w.r.t. $\langle \mathcal{H}, \Pi \rangle$.

We now consider a general causal diagram $\mathcal{G}$ where arbitrary latent endogenous variables $L$ exist and $Pa(\Pi) \subset O \setminus \{X, Y\}$. We will apply the latent projection [39], Def. 5) to transforms $G$ into a simplified causal diagram $\mathcal{H}$ discussed above. More specifically, we construct a causal diagram $\mathcal{G}'$ from $\mathcal{G}$ by marking each $V \in \mathcal{V} \setminus \{Pa(\Pi) \cup \{X, Y\}\}$ latent. We then apply Alg. 2 and obtain a simplified diagram $\mathcal{H} = \text{PROJECT}(\mathcal{G}')$ where $\mathcal{V} = O = Pa(\Pi) \cup \{X, Y\}$. Since PROJECT preserves topological relationships among observed nodes [39, Lem. 5], $(Y \not\perp X | Z)_{\mathcal{G}_X} \implies (Y \not\perp X | Z)_{\mathcal{H}_X}$. Following our previous argument, there exists a POSCM $M'$ associated with $\mathcal{H}$ such that for any policy $\pi \in \Pi$, $P(y | do(\pi); M') \neq P(y; M')$. By Lem. 3 we could construct a POSCM $M$ associated with $\mathcal{G}$ such that for any $\pi \in \Pi$, $P(y | do(\pi); M) = P(y | do(\pi); M')$ and $P(y; M) = P(y; M')$. It follows immediately that for any policy $\pi \in \Pi$, $P(y | do(\pi); M) \neq P(y; M)$, i.e., $P(y)$ is not imitable w.r.t. $\langle \mathcal{G}, \Pi \rangle$.

**Theorem 2** (Imitation by $\pi$-Backdoor). Given a causal diagram $\mathcal{G}$ and a policy space $\Pi$, $P(y)$ is imitable w.r.t. $\langle \mathcal{G}, \Pi \rangle$ if and only if there exists an $\pi$-backdoor admissible set $Z$ w.r.t. $\langle \mathcal{G}, \Pi \rangle$. Moreover, the imitating policy $\pi \in \Pi$ is given by $\pi(x | pa(\Pi)) = P(x | z)$.

**Proof.** We first prove the “if” direction. Given an $\pi$-backdoor admissible set $Z$ relative to $X, Y$ in $\mathcal{G}$. By Rule 2 of do-calculus, we have,

$$P(y) = \sum_{x, z} P(z)P(x | z)P(y | do(x), z).$$

Since $Z \subseteq Pa(\Pi)$, the conditional distribution $P(x | z)$ can be represented as a policy in $\Pi$. Let $\pi(x | Pa(\Pi)) = \pi(x | z) = P(x | z)$. We must have

$$P(y) = \sum_{x, z} P(z)\pi(x | z)P(y | do(x), z) = P(y | do(\pi)).$$

We now consider the “only if” direction. Suppose there exists no $\pi$-backdoor admissible set $Z$ relative to $(X, Y)$ in $\mathcal{G}$. We must have that $Z = An(y)_{\mathcal{G}} \cap Pa(\Pi)$ is not $\pi$-backdoor admissible, i.e., the independent relationship $(Y \not\perp X | Z)_{\mathcal{G}_X}$ does not hold. It follows immediately from Lem. 3 that $P(y) = 0$ is not imitable relative to $\langle \mathcal{G}, \Pi \rangle$.

**Lemma 2.** Given a causal diagram $\mathcal{G}$, a policy space $\Pi$, and an observational distribution $P(o)$, let $\langle S, \Pi' \rangle$ be an instrument w.r.t. $\langle \mathcal{G}, \Pi \rangle$. If $P(s)$ is $p$-imitable w.r.t. $\langle \mathcal{G}, \Pi', P(o) \rangle$, then $P(y)$ is
Let $S$ be a surrogate w.r.t. $\langle G, \Pi \rangle$. Following the proof of Lem. \ref{lem:identifiable},

\[
P(y|\text{do}(\pi); M) = \sum_s P(y|s, \text{do}(\pi))P(s|\text{do}(\pi))
= \sum_s P(y|s, \text{do}(x))P(s|\text{do}(\pi))
= \sum_s P(y|s)P(s|\text{do}(\pi)).
\]

Suppose $\langle S, \Pi' \rangle$ form an instrument w.r.t. $\langle G, \Pi \rangle$, i.e., $P(s|\text{do}(\pi))$ is identifiable w.r.t. $\langle G, \Pi' \rangle$. The above equation implies that $P(y|\text{do}(\pi))$ can be uniquely determined from $P(o, y)$ in any POSCM that induces $G$. That is, $P(y|\text{do}(\pi))$ is identifiable w.r.t. $\langle G \cup \{Y\}, \Pi \rangle$, which completes the proof. \hfill \Box

**Lemma 4.** Given a causal diagram $G$, a policy space $\Pi$, an observational distribution $P(o)$, and a subset $S \subseteq O$. $P(s)$ is p-imitable only if for any $S' \subseteq S$, $P(s')$ is p-imitable w.r.t. $\langle G, \Pi, P(o) \rangle$.

**Proof.** For any POSCM $M$ in $\mathcal{M}(G, P)$, if there exists a policy $\pi \in \Pi$ such that $P(s|\text{do}(\pi); M) = P(s; M)$, we must have for any $S' \subseteq S$,

\[
P(s'|\text{do}(\pi); M) = \sum_{s\setminus s'} P(s|\text{do}(\pi); M) = \sum_{s\setminus s'} P(s; M) = P(s; M).
\]

which completes the proof. \hfill \Box
Theorem 3. Given a causal diagram $G$, a policy space $\Pi$, and an observational distribution $P(o)$, if `IMITATE` returns a policy $\pi \in \Pi$, $P(y)$ is $p$-imitable w.r.t. $\langle G, \Pi, P(o) \rangle$. Moreover, $\pi$ is an imitating policy for $P(y)$ w.r.t. $\langle G, \Pi, P(o) \rangle$.

Proof. For a policy subspace $\Pi'$, Step 3 outputs a surrogate $S$ w.r.t. $\langle G, \Pi' \rangle$. Step 4 ensures that $\Pi'$ is an id-subspace w.r.t. $\langle G, \Pi, S \rangle$. That is, $\langle S, \Pi' \rangle$ forms an instrument w.r.t. $\langle G, \Pi \rangle$. Since `IMITATE` only outputs a policy $\pi$ when $P(s)$ is $p$-imitable w.r.t. $\langle G, \Pi', P(o) \rangle$, the statement follows from Lem. 2. \qed
B Causal Identification in POSCMs

In this section, we introduce algorithms for identifying causal effects in Partially Observable Structural Causal Models (POSCMs), which are sufficient and complete. We will consistently assume that $Y \subseteq O$, i.e., the primary outcomes $Y$ are all observed. For settings where $Y \nsubseteq O$, Corol. \[39, Alg. 5\] implies that $P(y|do(x))$ is always non-identifiable w.r.t. the causal diagram $G$. For convenience, we focus on the problem of determining whether the target effect is identifiable w.r.t. $G$. However, our algorithms could be easily extended to derive identification formulas of the causal effect.

We start with the identification of causal effects induced by atomic interventions $do(x)$. Formally,

**Definition 9 (Identifiability (Atomic Interventions)).** Given a causal diagram $G$, let $Y$ be an arbitrary subset of $O$. $P(y|do(x))$ is said to be identifiable w.r.t. $G$ if $P(y|do(x); M)$ is uniquely computable from $P(o; M)$ and $\pi$ for any POSCM $M \in \mathcal{M}(G)$.

A causal diagram $G$ is said to be semi-Markovian if it does not contain any latent endogenous variables $L$; or equivalently, $V = O$. \[39, Alg. 5\] is an algorithm that identify causal effects, say $P(y|do(x))$, from the observational distribution $P(o)$ in a semi-Markovian diagram $G$, which we consistently refer to as $\text{IDENTIFYHELPER}$. More specifically, $\text{IDENTIFYHELPER}$ takes as input a set of action variables $X \subseteq O$, a set of outcome variables $Y \subseteq O$ and a semi-Markovian causal diagram $G$. If $P(y|do(x))$ is identifiable w.r.t. $G$, $\text{IDENTIFYHELPER}$ returns an identification formula that represents $P(y|do(x))$ as an algebraic expression of the observational distribution $P(o)$; otherwise, $\text{IDENTIFYHELPER}$ returns “Fail”. \[13, 34\] showed that $\text{IDENTIFYHELPER}$ is complete from identifying effects from observational data with respect to semi-Markovian causal diagrams.

We will utilize $\text{IDENTIFYHELPER}$ to identifying causal effects in POSCMs where latent endogenous variables are allowed. The key to this reduction is an algorithm $\text{PROJECT} [39, Def. 5]$ that transforms an arbitrary causal diagram $G$ with observed endogenous variables $O$ into a semi-Markovian causal diagram $H$ such that its endogenous variables $V = O$. For completeness, we rephrase $\text{PROJECT}$ and describe it in Alg. \[3\].

---

**Algorithm 2: PROJECT \[39, Def. 5\]**

1: **Input:** A causal diagram $G$
2: **Output:** A causal diagram $H$ where all endogenous variables are observed, i.e., $V = O$
3: Let $O, L$ be, respectively, observed endogenous variables and latent endogenous variables in $G$
4: Let $H$ be a causal diagram constructed as follows.
5: for each observed $V \in O$ in $G$ do
6:  Add an observed node $V$ in $H$
7: end for
8: for each pair $S, E \in O$ in $G$ s.t. $S \neq E$ do
9:  if there exists a directed path $S \to E$ in $G$ then
10:    Add a directed edge $S \to E$ in $H$
11:  else if there exists a path $S \to V_1 \to \cdots \to V_n \to E$ in $G$ s.t. $V_1, \cdots V_n \in L$ then
12:    Add a directed edge $S \to E$ in $H$
13:  else if there exists a bidirected edge $S \leftrightarrow E$ in $G$ then
14:    Add a bidirected edge $S \leftrightarrow E$ in $H$
15:  else if there exists a path $S \leftarrow V_{i,1} \leftarrow \cdots \leftarrow V_{i,n} \leftarrow V_{r,m} \to \cdots \to V_{r,1} \to E$ in $G$ s.t. $V_{i,1}, \cdots, V_{i,n}, V_{r,1}, \cdots, V_{r,m} \in L$ then
16:    Add a bidirected edge $S \leftrightarrow E$ in $H$
17:  else if there exists a path $S \leftarrow V_{i,1} \leftarrow \cdots \leftarrow V_{i,n} \leftarrow V_c \to V_{r,m} \to \cdots \to V_{r,1} \to E$ in $G$ s.t. $V_{i,1}, \cdots, V_{i,n}, V_c, V_{r,1}, \cdots, V_{r,m} \in L$ then
18:    Add a bidirected edge $S \leftrightarrow E$ in $H$
19: end if
20: end for
21: Return $H$.

---

\[6\] In the original text, the semi-Markovian causal diagram $G$ is implicitly assumed; \[39, Alg. 5\] only takes $X, Y$ as input. We rephrase the algorithm and explicitly represent the dependency on the causal diagram $G$.  

---

17
The following lemma, introduced in [19 Props. 2-3], shows that the parameter space of observational distributions $P(a)$ and interventional distributions $P(y|do(x))$ induced by POSCMs associated with a causal diagram $G$ is always equivalent to that induced by the corresponding semi-Markovian causal diagram $H = \text{PROJECT}(G)$.

**Lemma 6.** Given a causal diagram $G$, let $H = \text{PROJECT}(G)$. For any POSCM $M_i$, associated with $G$, there exists a POSCM $M_2$ associated with $H$ such that $P(y|do(x); M_1) = P(y|do(x); M_2)$ for any $X, Y \subseteq O$, and vice versa.

**Proof.** The statement follows from [19 Props. 2-3].

Lem. 6 implies a general algorithm for identifying $P(y|do(x))$ in a causal diagram $G$ with latent endogenous variables: it is sufficient to consider the identifiability of $P(y|do(x))$ in the projection $H = \text{PROJECT}(G)$. We describe such an algorithm in Alg. 3.

**Algorithm 3: IDENTIFY (Atomic Interventions)**

1: **Input:** A causal diagram $G$, primary outcomes $Y \subseteq O$.
2: Let $H = \text{PROJECT}(G)$.
3: if IDENTIFYHELPER($\{X\}, Y, H$) = FAIL then
4: Return No.
5: else
6: Return Yes.
7: end if

**Corollary 4.** Given a causal diagram $G$, let $Y$ be an arbitrary subset of $O$. IDENTIFY($G, Y$) = “Yes” if and only if $P(y|do(x))$ is identifiable w.r.t. $G$.

**Proof.** Lem. 6 implies that $P(y|do(x))$ is identifiable w.r.t. a causal diagram $G$ if and only if $P(y|do(x))$ is identifiable w.r.t. the projection $H = \text{PROJECT}(G)$. To see this, suppose $P(y|do(x))$ is not identifiable w.r.t. $G$. That is, there exist two POSCMs $M_1, M_2$ associated with $G$ such that $P(\alpha; M_1) = P(\alpha; M_2)$ while $P(y|do(x); M_1) \neq P(y|do(x); M_2)$. By Lem. 6 for any $M_i$ and $i = 1, 2$, we could find a POSCM $M'_i$ associated with $H$ such that $P(\alpha; M'_i) = P(\alpha; M_i)$ and $P(y|do(x); M'_i) = P(y|do(x); M_i)$. This implies that we could obtain two POSCMs $M'_1, M'_2$ associated with $H$ such that $P(\alpha; M'_1) = P(\alpha; M'_2)$ while $P(y|do(x); M'_1) \neq P(y|do(x); M'_2)$, i.e., $P(y|do(x))$ is identifiable w.r.t. $H$. Similarly, we could prove the “only if” direction.

Since IDENTIFY returns “Yes” if and only if IDENTIFYHELPER finds an identification formula of $P(y|do(x))$ and IDENTIFYHELPER is sound and complete, the statement is entailed.

**B.1 Identifying Conditional Plans**

We will next study the general problem of identifying causal effects $P(y|do(\pi))$ induced by interventions $do(\pi)$ following conditional plans in a policy space $\Pi$. The formal definition of such identifiability is given in Def. 2. Similar to atomic interventions, we consider first a simpler setting where the causal diagram $G$ is semi-Markovian, without latent endogenous variables. Given a causal diagram $G$, we denote by $G_{\Pi}$ a manipulated diagram obtained from a subgraph $G_{\pi}$ by adding arrows from nodes in $Pa(\Pi)$ to the action node $X$. Let a set $Z = an(Y)_{\Pi} \setminus \{X\}$. Following Eq. 15, the interventional distribution $P(y|do(\pi))$ could be written as follows:

$$P(y|do(\pi)) = \sum_{x,z} P(y, z|do(x), do(v \setminus (y \cup z \cup \{x\})))\pi(x|pa(\Pi))$$

$$= \sum_{x,z} P(y, z|do(x))\pi(x|pa(\Pi)).$$

Among the above equations, the last step follows from Rule 3 of do-calculus [29 Thm. 3.4.1]. More specifically, if there is a directed path from a node $V_i \in V \setminus (Y \cup Z \cup \{X\})$ to a node in $Y, Z$ in the subgraph $G_{\pi}$, $V_i$ must also be included in set $Z$. That is, $(Y, Z \perp V \setminus (Y \cup Z \cup \{X\}))_{\Phi_{\pi}(V \setminus (Y \cup Z \cup \{X\}))}$, which implies $P(y, z|do(x), do(v \setminus (y \cup z \cup \{x\})) \perp \{\alpha\})$.
\{x\}) = P(y, z|do(x)). It is thus sufficient to identify the causal effect \(P(y, z|do(x))\) induced by atomic intervention \(do(x)\) w.r.t. \(G\). It showed that such an algorithm is complete for identifying \(P(y|do(\pi))\) in a semi-Markovian causal diagram \(G\) where all endogenous variables are observed.

**Lemma 7.** Given a causal diagram \(G\) and a policy space \(\Pi\), let \(Y\) be an arbitrary subset of \(O\). Assume that \(G\) is semi-Markovian, i.e., \(V = O\). Let \(Z = an(Y)_G \setminus \{X\}\). \(P(y|do(\pi))\) is identifiable w.r.t. \(\langle G, \Pi \rangle\) if and only if \(P(y, z|do(x))\) is identifiable w.r.t. \(G\).

**Proof.** The statement follows immediately from [5] Corol. 2. \qed

We are now ready to consider the identification of \(P(y|do(\pi))\) w.r.t. a policy space \(\Pi\) and a general causal diagram \(G\) where latent endogenous variables \(L\) are present. Our next result shows that it is sufficient to identify \(P(y|do(\pi))\) in the corresponding semi-Markovian projection \(H = \text{PROJECT}(G)\).

**Lemma 8.** Given a causal diagram \(G\) and a policy space \(\Pi\), let \(H = \text{PROJECT}(G)\). For any POSCM \(M_1\) associated with \(G\), there exists a POSCM \(M_2\) associated with \(H\) such that \(P(y|do(\pi); M_1) = P(y|do(\pi); M_2)\) for any \(\pi \in \Pi\). Let \(Y \subseteq O\), and vice versa.

**Proof.** For any policy \(\pi \in \Pi\), \(P(y|do(\pi))\) could be written as a function of \(P(y, z|do(x))\) and \(\pi\) following Eq. (3). Therefore, the statement is implied by Lem. 6. \qed

**Algorithm 4: IDENTIFY**

1: Input: A causal diagram \(G\), a policy space \(\Pi\), primary outcomes \(Y \subseteq O\).
2: Let \(H = \text{PROJECT}(G)\).
3: Let \(Z = \text{an}(Y)_H \setminus \{X\}\).
4: if IDENTIFYHELPER(\(\{X\}, Y \cup Z, H\)) = FAIL then
5: Return NO.
6: else
7: Return YES.
8: end if

Details of our algorithm IDENTIFY is described in Alg. 4. It takes as input a causal diagram \(G\), a policy space \(\Pi\) and a set of observed outcomes \(Y\). At Step 1, it obtains a projection \(H\) of \(G\) such that all endogenous variables \(V\) in \(H\) are observed. It then constructs the covariates \(Z\) following Lem. 7 (Step 3). Finally, IDENTIFY calls IDENTIFYHELPER to identify \(P(y, z|do(x))\) in the projection \(H\). It outputs “NO” if IDENTIFYHELPER fails to find an identification formula of \(P(y, z|do(x))\); otherwise, it outputs “YES”. Since Lem. 8 shows that the parameter space of \(P(y|do(\pi))\) and \(P(o)\) induced by POSCMs \(M \in \mathcal{M}(G)\) is equivalent to that induced by instances in the family \(\mathcal{M}(H)\) of projection \(H\), the soundness and completeness of IDENTIFY is entailed.

**Corollary 5.** Given a causal diagram \(G\) and a policy space \(\Pi\), let \(Y\) be an arbitrary subset of \(O\). IDENTIFY(\(G, \Pi, Y\)) = YES if and only if \(P(y|do(\pi))\) is identifiable w.r.t. \(\langle G, \Pi \rangle\).

**Proof.** Lem. 8 implies that \(P(y|do(\pi))\) is identifiable w.r.t. \(\langle G, \Pi \rangle\) if and only if \(P(y|do(\pi))\) is identifiable w.r.t. \(\langle H, \Pi \rangle\) where \(H = \text{PROJECT}(G)\). The proof is similar to Corol. 4. It follows from Eq. (3) and Lem. 7 that IDENTIFYHELPER does not fail if and only if \(P(y|do(\pi))\) is identifiable w.r.t. \(\langle H, \Pi \rangle\). The soundness and completeness of IDENTIFY is thus entailed. \qed
C  \textbf{LISTIdSPACE}

In this section, we describe Algorithm \textsc{LISTIdSPACE} that finds all identifiable subspaces with respect to a causal diagram \( \mathcal{G} \), a policy space \( \Pi \) and a set of observed variables \( Y \subseteq O \). The details of \textsc{LISTIdSPACE} are shown in Alg. 6

\begin{algorithm}[H]
\caption{\textsc{LISTIdSPACE}}
\begin{algorithmic}[1]
\STATE \textbf{Input:} \( \mathcal{G}, \Pi, Y \).
\STATE \textsc{LISTIdSPACEHelper}(\( \mathcal{G}, Y, \{ \pi : \mathcal{P}_X \}, \Pi \)).
\end{algorithmic}
\end{algorithm}

\begin{algorithm}[H]
\caption{\textsc{LISTIdSPACEHelper}}
\begin{algorithmic}[1]
\STATE \textbf{Input:} \( \mathcal{G}, Y \) and policy spaces \( \Pi_L, \Pi_R \) such that \( \Pi_L \subseteq \Pi_R \).
\IF {\textsc{IDENTIFY}(\( \mathcal{G}, \Pi_L, Y \)) = \text{YES}}
\STATE If \( L = R \) then Output \( \Pi_L \).
\ELSE
\STATE Pick an arbitrary \( V \in P_a(\Pi_R) \setminus P_a(\Pi_L) \).
\STATE \textsc{LISTIdSPACEHelper}(\( \mathcal{G}, Y, \Pi_L \cup \{V\}, \Pi_R \)).
\STATE \textsc{LISTIdSPACEHelper}(\( \mathcal{G}, Y, \Pi_R \setminus \{V\} \)).
\ENDIF
\end{algorithmic}
\end{algorithm}

It calls a subroutine \textsc{LISTIdSPACEHelper} which takes as input the diagram \( \mathcal{G} \), variables \( Y \) and two policy subspace \( \Pi_L, \Pi_R \) of \( \Pi \) such that \( \Pi_L \subseteq \Pi_R \). More specifically, \textsc{LISTIdSPACEHelper} performs backtrack search to enumerate identifiable subspaces \( \Pi' \) w.r.t. \( \langle \mathcal{G}, \Pi, Y \rangle \) such that \( \Pi_L \subseteq \Pi' \subseteq \Pi_R \). It aborts branches that will not lead to such identifiable subspaces. The aborting criterion (Step 2 of Alg. 6) follows the observation that \( P(\pi|\text{do}(\pi)) \) is identifiable w.r.t. a policy space \( \Pi \) only if it is identifiable w.r.t. any subspace \( \Pi' \subseteq \Pi \). At Step 5, it picks an arbitrary variable \( V \) that is included in the input covariates of \( \Pi_R \) but not in \( \Pi_L \). Let \( \Pi \cup \{V\} \) denote a policy space obtained from \( \Pi \) by including \( V \) as part of the input covariates, i.e., \( \Pi \cup \{V\} = \{ \pi : \mathcal{P}_{Pa(\Pi),V} \mapsto \mathcal{P}_X \} \). Similarly, we define \( \Pi \cup \{V\} = \{ \pi : \mathcal{P}_{Pa(\Pi),V} \mapsto \mathcal{P}_X \} \). \textsc{LISTIdSPACEHelper} then recursively returns all identifiable subspaces \( \Pi' \) w.r.t. \( \langle \mathcal{G}, \Pi, Y \rangle \); the first recursive call returns i-subspaces taking \( V \) as an input and the second call return all i-subspaces that does not consider \( V \).

\textbf{Theorem 4.} Given a causal diagram \( \mathcal{G} \), a policy space \( \Pi \), and an oracle access to \textsc{IDENTIFY}, let \( Y \) be an arbitrary subset of \( O \). \textsc{LISTIdSPACE}(\( \mathcal{G}, X, Y, \emptyset, Pa(\Pi) \)) enumerates identifiable subspaces w.r.t. \( \langle \mathcal{G}, \Pi, Y \rangle \) with polynomial delay \( \mathcal{O}(|Pa(\Pi)|) \).

\textbf{Proof.} The recursive calls at Steps 6 and 7 guarantees that \textsc{LISTIdSPACEHelper} generates every i-subspaces \( \Pi' \) exactly once. Since every leaf will output an i-subspace, the tree height is at most \( |Pa(\Pi)| \) and the existence check is performed by \textsc{IDENTIFY} oracle, the delay time is \( \mathcal{O}(|Pa(\Pi)|) \). \( \square \)
D Experiments

We perform 4 experiments, each designed to test different aspects of causal imitation learning. The basic results of the last two experiments are summarized in the paper’s main text.

1. **Data-Dependent Binary Imitability** - given a binary model, we show that by sampling uniformly from binary distributions satisfying the graph in figure 1c (called the “front-door”), naïve imitation results in a biased answer. We also observe that in binary frontdoor models, 50% of distributions exhibit data-dependent imitability, despite not being imitable in generality.

2. **GAN-Based Binary Imitability** - Generative Adversarial Networks [9] allow explicit parameterization of a model, and can therefore be used to imitate causal mechanisms in the presence of confounding. We show in various binary models that a GAN-based approach to imitation leads to positive results, both with standard and data-dependent imitation.

3. **Highway Driving** - The binary models in experiment 2 show that GANs function in the causal imitation setting, however, one can simply use the corresponding formulae to efficiently get answers. In this experiment, we use the HighD dataset to make two variables continuous, and to show that naïvely choosing information to use for imitation can lead to bias.

4. **MNIST Digits** - The final experiment shows that the GAN-based approach is capable of handling complex, high-dimensional probability distributions. To show this, we perform data-dependent imitation on a frontdoor graph, replacing a node with pictures of MNIST digits, to represent a complex, high-dimensional probability distribution.

Experiments 3 and 4 are reported in the main text. However, each experiment builds upon the ideas introduced in the preceding experiment, so it is recommended that readers go in order. Details of each experiment are described in its own subsection. For an discrete variable $X$, we will consistently use $x_i$ to represent an assignment $X = i$; therefore, we write $P(y_1|\text{do}(x_0)) = P(Y = 1|\text{do}(X = 0))$.

### D.1 Data-Dependent Binary Imitability

If all variables in a causal model are discrete, one can find the imitating policy through the solution of a series of linear systems. As an example, in the front-door graph (Fig. 1c and shown below), with all variables binary, and given an observational distribution $P(x, w, s)$ that is amenable to data-dependent imitation, the imitating policy $\pi$ for $X$ has probability of $\pi(x_1) = \alpha$:

$$\alpha P(s|\text{do}(x_1)) + (1 - \alpha) P(s|\text{do}(x_0)) = P(s)$$

$$\Rightarrow \alpha = \frac{P(s) - P(s|\text{do}(x_0))}{P(s|\text{do}(x_1)) - P(s|\text{do}(x = 0))} \quad (4)$$

In this experiment, we show the bias of cloning $\pi(x) = P(x)$ as compared to imitation using Eq. (4) in binary models. It is important to note that we limited this experiment to binary models to permit the computation of optimal policies explicitly - higher dimensional/continuous models are likely to show different average bias due to their extra degrees of freedom.

We generate $1 \times 10^5$ random instances where $X, W, S$ are binary variables. Probability distributions consistent with the graph decompose as $P(x, w, s, y) = P(x)P(w|x)P(s|x, w)P(y|s)$, so we draw uniformly over $[0, 1]$ for each conditional probability (i.e. $P(x), P(w|x), P(s|x, w), P(y|s) \sim U(0, 1)$). We report in Fig. 6 the L1 distance between the interventional distribution $P(y|\text{do}(\pi))$ induced by the imitator ($ci$ and $bc$) and the actual expert’s reward distribution $P(s)$ over $1 \times 10^5$ generated instances. The causal imitation learning approach uses Eq. (4), while naïve cloning directly imitates $P(X)$, both using sample averages. We find that the causal imitation ($ci$, average L1 = 0.0016) dominates the naïve cloning approach ($bc$, average L1 = 0.0147). More interestingly, we find 50% of generated instances are p-imitable; this suggests that leveraging the observational distribution is beneficial in many imitation learning settings.
When the models are sampled uniformly from probability distributions compatible with Fig. 1c, and for which there exists a perfect policy for imitating $P(y)$, a naïve cloning of $\pi(x) = P(x)$ (a) yields a biased answer, while our approach (b) gets the answer with only sampling error.

D.2 GAN-Based Binary Imitability

Given an identification formula for the causal effect of a target variable $X$ on $S$, with $X$ conditioned on a mediator $W$, there is a separate system of equations for each instantiation of the set $W$, akin to the one shown in Eq. (4). This means that the number of systems of equations to solve is exponential in the size of $W$, and can’t easily be approached when the model contains continuous or high-dimensional variables.

To avoid reliance on these adjustment formulae, we solve for imitating policies through direct modeling of the SCM [21, 10]. In particular, we follow a similar procedure to [16, 11], by training a generative adversarial network (GAN) to imitate the observational distribution, with a separate generator for each observable variable in the causal graph.

The advantage of this approach is that once a model is trained that faithfully reproduces the observational distribution on a given causal graph, any identifiable quantity will be identical in the trained model as in the original distribution, no matter the form of the underlying mechanisms and latent variables [29, Definition 3.2.4]. This means that so long as you have an instrument for $X$, one can use the trained generator to optimize for a conditional policy by directly implementing it in the model, as shown in Lem. 2.

In this experiment, we show that such an approach is, in fact, practical. To maintain the ability to compare our results with the ground truth, we use binary variables for each node in a tested graph. This restriction is loosened in the final two experiments - our goal here is to show that it is possible to get very accurate model reconstruction and imitation (with accurate intervention effects) with a GAN, even when there are latent variables.

Like in the previous experiment, for “ground-truth” data, our models are sampled uniformly from the space of factorized distributions. For example, for graph 1 in Table 1, one can factorize $P(x, y, z) = P(x)P(z|x)P(y|x, z)$, and can choose ground-truth probabilities $P(x), P(z|x'), P(y|x', z') \sim U(0, 1), \forall x, y, z, x', z'$.

For the GANs, we adapt discrete BGAN [12] to arbitrary SCM. The advantage of f-divergence-based approaches (such as BGAN) is that the f-GAN discriminators explicitly optimize a lower bound on the value of the chosen f-divergence. In particular, defining an f-divergence over two distributions $P$ and $Q$ as $D_f(P||Q) = \int_X q(x)f\left(\frac{p(x)}{q(x)}\right)dx$ with $f^*$ as the convex conjugate of $f$, and using $T$ as the set of functions that a neural network can implement, the f-GAN discriminator optimizes the following [27, 26]:

$$D_f(P||Q) \geq \sup_{T \in \mathcal{T}} (\mathbb{E}_{x \sim P}[T(x)] - \mathbb{E}_{z \sim Q}[f^*(T(x))])$$

with a tight bound for $T^*(x) = f'\left(\frac{p(x)}{q(x)}\right)$. This means that by choosing an appropriate function $f$ and the class of functions $T$, one can estimate the divergence through optimization over $T$ [26].

The KL divergence is an example of an f-divergence with $f(x) = x \log(x)$.
Table 1: Results of imitation using GANs in randomly sampled distributions. Each number is averaged over 100 randomly sampled models. Experiments in Row 3 and 4 are run with p-imitable, and not p-imitable distributions, respectively.

| # | Graph | \(|y_{x_0} - \hat{y}_{x_0}|\) | \(|y_{x_1} - \hat{y}_{x_1}|\) | \(|\alpha - \hat{\alpha}|\) | \(|y - \hat{y}|\) |
|---|---|---|---|---|---|
| 1 | ![Graph 1](image1.png) | 0.0063 | 0.0024 | 0.0655 | 0.0016 |
| 2 | ![Graph 2](image2.png) | 0.0017 | 0.0044 | 0.0942 | 0.0059 |
| 3 | ![Graph 3](image3.png) | 0.0018 | 0.0022 | 0.0788 | 0.0030 |
| 4 | ![Graph 4](image4.png) | 0.0025 | 0.0027 | 0.2132 | 0.0206 |

Table 1: Results of imitation using GANs in randomly sampled distributions. Each number is averaged over 100 randomly sampled models. Experiments in Row 3 and 4 are run with p-imitable, and not p-imitable distributions, respectively.

this trained discriminator \(T(x)\), one can update generator \(Q_\theta\) to be more similar to \(P\) with a update similar to a policy-gradient [2].

To adapt these GANs to work with causal models, we create a separate generator for each variable \(V \in V\), which is given as input samples from \(P_{aV}\), and outputs multinomial probabilities. Each latent variable is explicitly sampled from \(N(0, 1)^k\) (with \(k = 3\) in these experiments), and given as input to its children. Finally, instead of a single discriminator for the entire model, we exploit independence relations in the graph to localize optimization. In particular, in a Markovian causal model (without latent variables), one can create a separate discriminator for each variable, to directly learn the conditional distribution for each node. However, once there are latent variables, this variable-focused approach no longer captures all dependencies. Instead, we construct a separate conditional discriminator for each set of nodes that have a path made up of entirely bi-directed edges (c-components [42]), which allows each discriminator to specialize to a part of the model.

As an example, for the graph 1 in Table 1, the generator for \(X\) gets as input a sample \(u \sim N(0, 1)^3\), and outputs a probability of 1. The generator for \(Z\) gets as input a sample according to the probability of \(X\), and outputs a conditional probability of \(Z\). Finally, \(Y\) gets as input \(u\) and the sampled value of \(Z\). This graph has 2 c-components (\(\{X, Y\}\) and \(\{Z\}\)), which corresponds to discriminators \(P(z|x)\) and \(P(x, y|z)\).

Similarly, once the model is optimized, the policy \(\pi\) is trained by manually replacing the generator for \(X\) with a new, untrained generator for \(\pi\), and training it as a GAN with a discriminator comparing samples of \(Y\) in the original model (i.e., \(Y \sim P(y)\)) with samples of \(Y\) in the intervened model (i.e., \(Y \sim P(y|do(\pi); M)\) where \(M\) is the parametrized model learned by generators.)

The graphs in Table 1 were each chosen to demonstrate a different aspect of imitability. The first graph is imitable by direct parents (Thm. 1), and corresponds to existing approaches to imitation, where an agent gets observations identical to the expert. The second graph demonstrates an example where an expert has additional information from a latent variable, but the agent can still successfully imitate \(P(y)\) by using information that is not used by the expert, i.e., the \(\pi\)-backdoor admissible set \(\{W\}\) (Thm. 2). Finally, Graphs 3 and 4 demonstrate data-dependent imitability (Def. 5). Graph 3 focuses on distributions that are amenable to imitation, meaning that imitation of \(P(y)\) is possible without knowledge of the latent variable. Graph 4 uses only non-imitable distributions. Sampled uniformly, half of these instances are p-imitable, and half are not, so the expected performance of an unknown distribution is the average of Row 3 and 4.

The table columns show the average L1 distance between the computed interventional distributions, the predicted & optimal policy, as well as the effect \(P(y|do(\pi))\) of the learned policy vs the observed \(P(y)\). Each element of the table shows an average value of its corresponding distance from ground-truth over 100 runs (with each run sampling a different ground-truth probability distribution over the graph), overlaid over the histogram of distances over the 100 distributions/trained GANs.
We demonstrate this in a two-step procedure. First, we performed an automated adversarial search over the interventional distribution of atomic intervention $\text{do}(x)$. The $|y - \hat{y}|$ represents the differences between the ground-truth $E[y|\text{do}(x)]$ and the imitated $E[y|\text{do}(x); M]$ in the parametrized model $M$. Notice that the policies are often conditioned, which is not reflected in these values.

The $|y - \hat{y}|$ value represents the difference in the ground-truth expert’s reward $E[y]$ with the distribution $E[y|\text{do}(\pi)]$ induced by the learned policy $\pi$ in the ground-truth model, meaning that the policy is trained in the imitated model, but is tested by replacing the true mechanism, as if the learned mechanism was tried in real life.

The main point of possible confusion could be $|\alpha - \hat{\alpha}|$ in the graphs where the policy is conditional. In the frontdoor cases, when the policy has no conditioning (Rows 3 and 4), the optimal value can be cleanly found, using Eq. (4). However, in the backdoor case (as seen in Row 1 and in Row 2), there can be multiple possible valid solutions. Here we show the precise procedure used to compute $|\alpha - \hat{\alpha}|$ in these two cases.

- In Row 1, which is commonly called the “backdoor” graph, we have defining $\alpha_1 = \pi(x_1|z_1)$ and $\alpha_0 = \pi(x_1|z_0)$:
  \[
P(y) = \sum_{x, z} P(z)\pi(x|z)P(y|x, z)
  = P(z_1)\alpha_1 P(y|x_1, z_1) + P(z_1)(1 - \alpha_1)P(y|x_0, z_1)
  + P(z_0)\alpha_0 P(y|x_1, z_0) + P(z_0)(1 - \alpha_0)P(y|x_0, z_0)
  = P(z_1)P(y|x_0, z_1) + P(z_0)P(y|x_0, z_0)
  + \alpha_1 P(z_1)(P(y|x_1, z_1) - P(y|x_0, z_1))
  + \alpha_0 P(z_0)(P(y|x_1, z_0) - P(y|x_0, z_0))
\]

This means that multiple possible $\alpha_0, \alpha_1$ satisfy the given constraint. In such situations, given $\hat{\alpha}_1$ and $\hat{\alpha}_0$ found by GAN, we find the “closest correct comparison” by minimizing

\[
P(z_0)|\hat{\alpha}_0 - \alpha_0| + P(z_1)|\hat{\alpha}_1 - \alpha_1|
\]

subject to the constraint in Eq. (5), and with $\alpha_1, \alpha_0 \in [0, 1]$. We then report this minimized value in the column labeled $|\alpha - \hat{\alpha}|$.

- Row 2 has more complex relations, since the policy has as inputs both values from $Z$ and $W$. However, the values of $W$ are irrelevant to imitating $y$, since $W$ is in a different c-component of the graph than $Y$. This means that we can use the same approach as for Row 1, considering only $Z$ (and averaging the policy over $W$ values).

### D.2.1 Discussion

The results in Table I suggest that GANs are capable of training accurate imitating policies in the presence of latent variables. Of particular note is the relatively large error in $\alpha$ can sometimes be present in the policies, despite the policies yielding very accurate samples of $y$. This happens when the imitable policy has $P(y|\text{do}(x_0))$ and $P(y|\text{do}(x_1))$ with very similar values, meaning that the policy has little effect on the probability of $Y$.

### D.3 Highway Driving

The purpose of this experiment is to demonstrate that when imitating $P(y)$, it is important to choose a set of covariates that is $\pi$-backdoor admissible (Def. 3). To witness, in Fig. 7a using knowledge of $W$ to imitate $X$ effectively adds confounding between $X$ and $Y$, possibly affecting the imitated distribution of $Y$. That is, $\{W\}$ is not $\pi$-backdoor admissible. Similarly, in Fig. 7b one must use knowledge of $Z$ when imitating $P(y)$, but using either only $W$ or both $Z$ and $W$ can lead to bias and inferior performance on reward measure $Y$. In other words, set $\{Z\}$ is $\pi$-backdoor admissible while $\{W, Z\}$ is not due to the active path $X \leftrightarrow L \rightarrow W \leftrightarrow Y$.

We demonstrate this in a two-step procedure. First, we performed an automated adversarial search over the binary distributions consistent with Fig. 7a to maximize $E[Y]$ when imitating just $P(x) = P(x)$, but minimizing $E[Y]$ when exploiting knowledge of $W$ to imitate $\pi(x|w) = P(x|w)$. This search led to the following mechanisms ($\oplus$ is xor):
Figure 7: The graphs used for the backdoor experiments. First, Fig. 7a was optimized adversarially with binary variables to maximize naïve imitation error, then continuous data from the highD dataset were added in Fig. 7b while maintaining the adversarial imitation error.

\[ P(l_1) = P(u_1) = \frac{\sqrt{5} - 1}{2} \approx 0.62 \quad W \leftarrow L \land U \quad X \leftarrow L \quad Y \leftarrow X \oplus U \]

This means that under “correct” imitation (not using knowledge of \( W \)), we have:

\[ \mathbb{E}[Y|\text{do}(\pi_1)] = \sum_{x,u} \mathbb{E}[Y|x,u] \pi_1(x) P(u) = 2\sqrt{5} - 4 \approx 0.472 \]

Under “incorrect” imitation (using \( W \)) we get:

\[ \pi_2(x|w) = P(x|w) = \frac{\sum_{l,u} P(u) P(l) P(x_1|l) P(w|l, u)}{\sum_{l,u} P(l) P(u) P(w|l, u)} = \begin{cases} 2\sqrt{5} - 2 \sqrt{5} - 1 & \approx 0.382 \text{ if } W = 0 \\ 1 & \text{ if } W = 1 \end{cases} \]

This means:

\[ \mathbb{E}[Y|\text{do}(\pi_2)] = \sum_{x,l,u,w} \mathbb{E}[Y|x,u] \pi_2(x|w) P(w|l,u) P(l) P(u) = 4\sqrt{5} - 9 \sqrt{5} - 3 \approx 0.2918 \]

Therefore, using the mechanisms above in Fig. 7a with binary variables, using \( W \) leads to a bias of 18%:

\[ \mathbb{E}[Y|\text{do}(\pi_1)] - \mathbb{E}[Y|\text{do}(\pi_2)] = 2\sqrt{5} - 4 - 4\sqrt{5} - 9 \sqrt{5} - 3 \approx 0.18 \]

Indeed, by sampling 10,000 data points, and using the empirical \( P(x) \) for one “imitator”, and \( P(x|w) \) for the other, we get 0.1793 difference between the two, corroborating this result.

### D.3.1 Making Things Continuous

We next show that this same issue can show up when variables are continuous. To achieve this, we adapted car velocity data from the highD dataset to a model similar to the binary version. In this model, one must use \( Z \) (velocity of the front car) to predict \( X \) (velocity of the driving car), but \( W \) would bias the prediction (with the same error as in previous section). The full model specification is as follows:

1. \( P(L = 1) = P(U = 1) = 0.62 \) (with values of \( L \) constructed from values of \( X \))
2. \( W \leftarrow L \land U \)
3. \( Z \) is velocity of preceding car from highD dataset.
4. \( X \) is velocity of current car from highD dataset. \( L \) was constructed such that \( X \) satisfies the relation \( L = I_{\{X - Z > -0.4\}} \) (this was achieved by choosing the threshold \(-0.4\) to give the correct distribution over \( L \)).
5. \[ Y \leftarrow U \land I_{\{X - Z \leq -0.4\}} \lor \neg U \land I_{\{X - Z > -0.4\}} \]

The values and mechanisms here were specifically chosen to have similar outputs to the previous model (Fig. 7a), despite using continuous car velocity data in place of boolean for node values. Fig. 7b shows a faithful graphical representation of the model. However, during the experiment, all algorithms are provided with only the causal diagram in Fig. 4a. That is, only independence relationships encoded in Fig. 4a are exploited.

![Graphical representation of the model](image)

Figure 8: Results of experiments using the continuous “backdoor” model. (a) and (c) are histograms of 100 independent runs of a direct supervised learning and GAN approach, respectively. (b) shows the distribution over X of a trained GAN model

We performed two separate imitation experiments, both of which show that (1) using the covariates set \{W, Z\} introduces biases into the imitator’s reward \( P(y|do(\pi)) \); (2) using the \( \pi\)-backdoor admissible set \{Z\} allows one to successfully imitate the expert’s performance \( P(y) \). The first experiment was imitation using standard supervised learning. Two separate (2 layer) neural networks were trained on a small subset of data using L2 loss over X. The first had as its input the values of \( Z \) (velocity of car ahead), while the second had as input values of \( W \) and \( Z \). The trained outputs of these networks were used for imitation using the full dataset, and the resulting policy was evaluated by its induced expected performance \( \mathbb{E}[Y|do(\pi)] \) (\( Y \) represents an unknown reward, so bigger is better). This experiment was repeated 100 times, giving a distribution over performance of trained models, shown in Fig. 8a. Note that while the performance varied widely, knowledge of \( W \) had a clear negative effect.

The outputs of a supervised learning algorithm are single values - they do not represent a distribution over possible values. This can affect imitation, since it is possible that the full range of the distribution is necessary for optimal imitation. We next trained two GANs to imitate the distribution over X, in the same way as done with the supervised models - one using only \( Z \), and one using both \( W \) and \( Z \). Due to the sampling needs of GANs, however, they had access to the full dataset over \( X, Z, W \).

This experiment was also repeated 100 times, giving the performance shown in Fig. 8c. Once again, the GAN using only \( Z \) has clearly superior performance. This difference exists despite both trained GANs seemingly recovering a good approximation over the distribution of \( X \), as seen in Fig. 8b.

D.4 MNIST Digits

The purpose of this experiment is to show how Algorithm Alg. using GANs for policy estimation could obtain reasonable imitation results in the p-imitable setting, even when the probability distribution of some observed endogenous variables is high-dimensional.

Specifically, we use the frontdoor graph, same as in the first experiment. There, when \( X \) and \( S \) are binary, we can compute the value \( \alpha \) using Eq. 5. However, in this case, the formula for the interventional distribution is [50]:

\[
P(s|do(\pi)) = \sum_x \pi(x) \sum_w P(w|x) \sum_{x'} P(s|x', w) P(x')
\]

Critically, computing the effect of an intervention here requires a sum (or integral if continuous) over \( W \). If \( W \) is a complex distribution or is high-dimensional, this quantity can be difficult to estimate.
Figure 10: In (a) and (b), results of a trained GAN conditional on $X = 0$ and $X = 1$ respectively (0 for FALSE, 1 for TRUE). In (c) is shown a histogram of the values obtained over multiple runs of the GAN. The red line represents naïve imitation, and the green line represents optimal imitation.

Once again, to allow us the ability to compare results to a ground-truth value, we repeat the procedure performed in the previous experiment, by first constructing a binary model with known characteristics, and then by replacing the binary value of $W$ with a high-dimensional distribution with a property that follows the same underlying mechanism as the original binary value. Specifically, we use the MNIST digits for 0 and 1, which are a $28 \times 28 = 784$ dimensional vector representing a complex probability distribution. The pictures of 0s replace “FALSE” values of $W$, while pictures of 1s replace “TRUE” values. This allows a direct translation between a binary variable and the desired complex distribution. The underlying binary distribution had the following mechanisms:

\[
\begin{align*}
P(u_1) &= 0.9 \\
P(x_1|u_0) &= 0.1 & P(x_1|u_1) &= 0.9 \\
P(w_1|x_1) &= 0.1 & P(w_1|x_0) &= 0.9 \\
P(s_1|u_0, w_0) &= 0.1 & P(s_1|u_0, w_1) &= 0.9 \\
P(s_1|u_1, w_0) &= 0.9 & P(s_1|u_1, w_1) &= 0.1
\end{align*}
\]

This leads to $P(s_1) = 0.245$, but with naïve cloning of $P(x)$, the resulting $P(s_1|\text{do}(\pi)) = 0.334$, a difference of 0.0888. We chose the reward signal $Y \leftarrow \neg S$, since the assumption is that the original values of $X$ were decided by an expert.

The GAN for the c-component $\{W\}$ is much larger than that of the other variables, since $W$ is an image, so we pre-trained this component, and inserted the trained version into the full graph for optimization. We repeated experiment 16 times, of which 3 runs were discarded due to collapse in the GANs associated with $P(w|x)$. The results are visible in Fig. 10. These results show that despite the high-dimensional nature of the distribution over $W$, the GAN was consistently able to perform better than naïve imitation, approaching the optimal value.