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, 30, 31, 32, 34, 34]. 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 [12, 13, 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 [35, 31, 24, 24, 24] and inverse reinforcement learning [25, 13, 26, 26]. The former focuses on learning a nominal expert policy that approximates the conditional distribution mapping 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

This example shows that even when one is able to perfectly mimic an optimal demonstrator, the fully observed [7], while another assumed that the primary outcome is observed (e.g., described above and our contributions. Some work considered settings in which the input to the expert policy is its dimension. We consistently use the abbreviation the domain of capital letters to denote random variables (1.1 Preliminaries

proofs in the complete technical report [15, Appendix A].

our results on high-dimensional, synthetic datasets. For the sake of space constraints, we provide all findings 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 data-generating process represented as a causal graph. (2) We develop a sufficient algorithm for determining the feasibility of imitation from demonstration data and qualitative knowledge about the space of probability distributions over {0, 1}; \( \oplus \) represents the exclusive-or operator. The expected reward \( \mathbb{E}[Y \mid \text{do}(\pi)] \) 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. Our task is to learn an imitating policy that achieves the expert’s performance from demonstration data in a structural causal model [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 [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 \( \mathbf{X} \), \(|X|\) 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.

\[1\]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 [7], while another assumed that the primary outcome is observed (e.g., Y in Fig. 1).
Calligraphic letters, e.g., $\mathcal{G}$, will be used to represent directed acyclic graphs (DAGs) (e.g., Fig. 1).

We denote by $\mathcal{G}_X$ the subgraph obtained from $\mathcal{G}$ by removing arrows coming into nodes in $X$; $\mathcal{G}_X$ is a subgraph of $\mathcal{G}$ 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 $\mathcal{G}$ is denoted by $\text{pa}(X)_\mathcal{G} = \cup_{x \in X} \text{pa}(X)_\mathcal{G}$. $\text{ch}$, $\text{de}$ and $\text{an}$ are similarly defined.

We write $\text{Pa}$, $\text{Ch}$, $\text{De}$, $\text{An}$ if arguments are included as well, e.g., $\text{De}(X)_\mathcal{G} = \text{de}(X)_\mathcal{G} \cup X$. A path from a node $X$ to a node $Y$ in $\mathcal{G}$ 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 $\mathcal{G}$, denoted by $(X \perp Y | Z)_\mathcal{G}$, if every edge path from nodes in one set to nodes in another are “blocked”. The criterion of blockage follows [29, Def. 1.2.3].

The basic semantic framework of our analysis rests on variables $\{\text{all possible policies}$ $P\}$ manipulated SCM of of $X$ independently following latent and expected value of a reward variable to decide the value of an action variable indicates the presence of an unobserved confounder (UC) affecting both latent variables, which is called the partially observable structural causal model (POSCM).

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 $\mathcal{G}$ (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., $U_{V_i} \cap U_{V_j} \neq \emptyset$.

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 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., $Y \in L \cap \text{De}(X)_\mathcal{G}$). A policy $\pi$ is a function mapping from values of covariates $P_{a^*} \subseteq O \setminus \text{De}(X)_\mathcal{G}$ to a probability distribution over $X$, which we denote by $\pi(x | P_{a^*})$. An intervention following a policy $\pi$, denoted by $\text{do}(\pi)$, 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 $\text{do}(\pi)$. Similar to atomic settings, the interventional distribution $P(v | \text{do}(\pi))$ is defined as the distribution over $V$ in the manipulated model $M_\pi$, given by,

$$P(v | \text{do}(\pi)) = \sum_u P(u) \prod_{v \in V \setminus \{X\}} P(v | P_{a^*}, u_v)\pi(x | P_{a^*}).$$ (1)

The expected reward of a policy $\pi$ is thus given by the causal effect $\mathbb{E}[Y | \text{do}(\pi)]$. The collection of all possible policies $\pi$ defines a policy space, denoted by $\Pi = \{\pi : \mathcal{P}_{a^*} \rightarrow \mathcal{P}_{X}\}$ (if $P_{a^*} = \emptyset$, $\Pi = \{\pi : \mathcal{P}_{X}\}$). For convenience, we define function $P_{a}(\Pi) = P_{a^*}$. A policy space $\Pi$ is a $^3$This definition will facilitate the more explicitly articulation of which endogenous variables are available to the demonstrator and corresponding policy at each point in time.

$^4$ $\mathcal{G}_{\mathcal{F}}$ is a causal diagram associated with the submodel $M_0$ induced by intervention $\text{do}(x)$.
which is far from the optimal expert reward, this is not the case:
for any policy associated with the causal diagram is not imitable w.r.t. \( P \) from the combinations of the observed data \( P(\mathbf{o}) \) 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 \( \mathcal{V} \). \( P(\mathbf{y}|\text{do}(\mathbf{\pi})) \) is said to be identifiable w.r.t. \( \langle \mathcal{G}, \Pi \rangle \) if \( P(\mathbf{y}|\text{do}(\mathbf{\pi}); M) \) is uniquely computable from \( P(\mathbf{\omega}; M) \) 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 \( \mathbb{E}[Y|\text{do}(\pi)] \):

**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., \( Y \cap \mathcal{L} \neq \emptyset \)), \( P(\mathbf{y}|\text{do}(\mathbf{\pi})) \) is not identifiable.

In other words, Corol. 1 shows that when the reward \( Y \) is latent, it is infeasible to uniquely determine values of \( \mathbb{E}[Y|\text{do}(\pi)] \) from \( P(\mathbf{\omega}) \). 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 \( Y \), \( \mathbb{E}[Y|\text{do}(\pi)] = \mathbb{E}[Y] \), 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 \( \mathcal{V} \). \( P(\mathbf{y}) \) is said to be imitable w.r.t. \( \langle \mathcal{G}, \Pi \rangle \) if there exists a policy \( \pi \in \Pi \) uniquely computable from \( P(\mathbf{\omega}) \) such that \( P(\mathbf{y}|\text{do}(\mathbf{\pi}); M) = P(\mathbf{y}; M) \) 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 \( P(\mathbf{\omega}) \) such that \( P(\mathbf{y}|\text{do}(\mathbf{\pi})) = P(\mathbf{y}|\text{do}(\mathbf{\pi}) \in M) \) in any POSCM \( M \) associated with the causal diagram \( \mathcal{G} \). Consider Fig. 3a as an example. \( P(\mathbf{y}) \) is imitable with policy \( \pi(x) = P(x) \) since by Eq. 1 and marginalization, \( P(\mathbf{y}|\text{do}(\mathbf{\pi})) = \sum_x P(\mathbf{y}|w)P(w|x)\pi(x) = \sum_x \sum_{w} P(\mathbf{y}|w)P(w|x)P(x) = P(\mathbf{y}) \). 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 \( \mathcal{V} \). \( P(\mathbf{y}) \) 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 \( P(\mathbf{\omega}; M_1) = P(\mathbf{\omega}; M_2) \) while there exists no policy \( \pi \in \Pi \) such that for \( i = 1, 2 \), \( P(\mathbf{y}|\text{do}(\pi); M_i) = P(\mathbf{y}; M_i) \).

It follows as a corollary that \( P(\mathbf{y}) \) is not imitable if there exists a POSCM \( M \) compatible with \( \mathcal{G} \) such that no policy \( \pi \in \Pi \) could ensure \( P(\mathbf{y}|\text{do}(\pi); M) = P(\mathbf{y}; M) \). For instance, consider the causal diagram \( \mathcal{G} \) and policy space \( \Pi \) in Fig. 3b. Here, the expert’s reward \( P(\mathbf{y}) \) 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, \( P(Y = 1|\text{do}(\pi)) = 0.5 \) 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 \( P(\mathbf{y}|\text{do}(\pi)) \) is identifiable, following the front-door criterion in \[29\] Thm. 3.3.4, but no

---

\[4\] The imitation is trivial if \( Y \not\in \text{De}(X)_{\mathcal{G}} \): by Rule 3 of \[29\] Thm. 3.4.1 (or \[6\] Thm. 1), \( P(\mathbf{y}|\text{do}(\pi)) = P(\mathbf{y}) \) for any policy \( \pi \). This paper aims to find a specific \( \pi \) satisfying \( P(\mathbf{y}|\text{do}(\pi)) = P(\mathbf{y}) \) even when \( Y \in \text{De}(X)_{\mathcal{G}} \).
policy imitates the corresponding $P(y)$. However, in some settings, the expert’s reward $P(y)$ is imitable but the imitator’s reward $P(y|\text{do}(\pi))$ cannot be uniquely determined. To witness, consider again the example in Fig. 1a. 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 $P(x|\text{pa}(\Pi))$.

Next, we consider the more challenging setting when policy spaces of the expert and learner disagree (i.e., the learner and expert have different views of the world, $f_X \not\subseteq \Pi$). We will leverage a graphical condition adopted from the celebrated backdoor criterion [29, Def. 3.3.1].

**Definition 4** ($\pi$-Backdoor). Given a causal diagram $\mathcal{G}$ and a policy space $\Pi$, a set $Z$ is said to satisfy the $\pi$-backdoor criterion w.r.t. $(\mathcal{G}, \Pi)$ if and only if $Z \subseteq \text{pa}(\Pi)$ and $(Y \perp \!\!\!\perp X|Z)_{\mathcal{G}_{\Delta}}$, which is called the $\pi$-backdoor admissible set w.r.t. $(\mathcal{G}, \Pi)$.

For concreteness, consider again the highway driving example in Fig. 1a. There exists no $\pi$-backdoor admissible set due to the path $X \leftarrow L \rightarrow Y$. Now consider a modified graph in Fig. 1b where edge $L \rightarrow Y$ is removed. $\{Z\}$ is $\pi$-backdoor admissible since $Z \subseteq \text{pa}(\Pi)$ and $(Y \perp \!\!\!\perp X|Z)_{\mathcal{G}_{\Delta}}$. Leveraging the imitation backdoor condition, our next theorem provides a full characterization for when imitating expert’s performance is achievable, despite the fact that the reward $Y$ is latent.

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

That is, one can learn an imitating policy from a policy space $\Pi' = \{\pi: \mathcal{G}_X \rightarrow \mathcal{P}_X\}$ that mimics the conditional probabilities $P(x|z)$ if and only if $Z$ is $\pi$-backdoor admissible. If that is the case, such a policy can be learned from data through standard density estimation methods. For instance, Thm. 2 ascertains that $P(y)$ in Fig. 1a is indeed non-imitable. On the other hand, $P(y)$ in Fig. 1b is imitable, guaranteed by the $\pi$-backdoor admissible set $\{Z\}$; the imitating policy is given by $\pi(x|z) = P(x|z)$.

### 3 Causal Imitation Learning with Data Dependency

One may surmise that the imitation boundary established by Thm. 2 suggests that when there exists no $\pi$-backdoor admissible set, it is infeasible to imitate the expert performance from observed trajectories of demonstrations. In this section, we will circumvent this issue by exploiting actual parameters of the observational distribution $P(o)$. In particular, we denote by $\mathcal{M}_{(\mathcal{G}, P)}$ a subfamily of candidate models in $\mathcal{M}_{(\mathcal{G})}$ that induce both the causal diagram $\mathcal{G}$ and the observational distribution $P(o)$, i.e., $\mathcal{M}_{(\mathcal{G}, P)} = \{M \in \mathcal{M}_{(\mathcal{G})}: P(o; M) = P(o)\}$. We introduce a refined notion of imitability that will explore the quantitative knowledge of observations $P(o)$ (to be exemplified). Formally,

**Definition 5** (Practical Imitability). 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, $\pi$-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. $(\mathcal{G}, \Pi, P(o))$, then $P(y)$ is $\pi$-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 $\mathcal{G}$. In other words, for an expert’s performance $P(y)$ that is non-imitable w.r.t. $(\mathcal{G}, \Pi)$, it could still be $p$-imitable after analyzing actual probabilities of the observational distribution $P(o)$.

For concreteness, consider again $P(y)$ in Fig. 1d 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: $X \leftarrow U_X \oplus U_Y$, $W \leftarrow X \oplus U_W$, $Y \leftarrow W \oplus U_Y$ 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 $P(y|do(x)) = \sum_w P(w|x) \sum_{x'} P(y|w, x') P(x')$ [29 Thm. 3.3.4]. We thus have $P(Y = 1|do(X = 0)) = 0.82$ which coincides with $P(Y = 1) = 0.82$, i.e., $P(y)$ is $p$-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 $\mathcal{O}$.

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 $\mathcal{G}$, a policy space $\Pi$, let $S$ be an arbitrary subset of $O$. $S$ is an imitation surrogate (for short, surrogate) w.r.t. $(\mathcal{G}, \Pi)$ if $(Y \perp \!\!\!\perp \hat{X}|S)_{\mathcal{G} \cup \Pi}$ where $\mathcal{G} \cup \Pi$ is a supergraph of $\mathcal{G}$ by adding arrows from $Pa(\Pi)$ to $X$; $\hat{X}$ 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. $(\mathcal{G}, \Pi)$. Consider as an example Fig. 1d where the supergraph $\mathcal{G} \cup \Pi$ coincides with the causal diagram $\mathcal{G}$. 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 $P(y|do(\pi)) = \sum_{s, w} P(y|s) P(s|w) P(w|x) \pi(x) P(u) = \sum_s P(y|s) P(s|do(\pi))$. 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 resultant 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. $(\mathcal{G}, \Pi)$. For example, in Fig. 1d $S$ is a surrogate w.r.t. $(\mathcal{G}, \Pi)$, but $P(s|do(\pi))$ is not identifiable due to collider $Z (\pi$ 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 $\mathcal{G}$, 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. $(\mathcal{G}, \Pi, S)$ if $P(s|do(\pi))$ is identifiable w.r.t. $(\mathcal{G}, \Pi')$.

Consider a policy subspace $\Pi' = \{\pi : \mathcal{P}_X\}$ described in Fig. 1d (i.e. $\pi$ that does not exploit information from covariates $Z$). $P(s|do(\pi))$ is identifiable w.r.t. $(\mathcal{G}, \Pi')$ 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, $\{S\}$ and $\Pi'$ forms an instrument that allows one to solve the imitation learning problem in Fig. 1d.

**Definition 8 (Imitation Instrument).** Given a causal diagram $\mathcal{G}$ 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. $(\mathcal{G}, \Pi')$ and $\Pi'$ is an id-subspace w.r.t. $(\mathcal{G}, \Pi, S)$.

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

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 $P(s|do(\pi)) = P(s)$.

---

5Consider a variation of the model where $P(U_W = 1) = 0.7$. $P(y)$ is $p$-imitable with $\pi(X = 0) = 0.75$. 

6
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 \((\mathcal{G}, \Pi, P(o))\). 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. \((\mathcal{G}, \Pi', P(o))\). 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 \(\mathcal{G} \cup \{Y\}\) denote a causal diagram obtained from \(\mathcal{G}\) by making reward \(Y\) observed. The following proposition suggests that it suffices to consider only identifiable subspaces w.r.t. \((\mathcal{G} \cup \{Y\}, \Pi, Y)\).

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

Our algorithm \textsc{Imitate} is described in Alg. 1. We assume access to an \textsc{Identify} oracle \cite{31, 33} that takes as input a causal diagram \(\mathcal{G}\), a policy space \(\Pi\) and a set of observed variables \(S\). If \(P(s|do(\pi))\) is identifiable w.r.t. \((\mathcal{G}, \Pi)\), \textsc{Identify} returns “YES”; otherwise, it returns “NO”. For details about the \textsc{Identify} oracle, we refer readers to \cite{15} Appendix B]. More specifically, \textsc{Imitate} takes as input a causal diagram \(\mathcal{G}\), a policy space \(\Pi\) and an observational distribution \(P(o)\). At Step 2, \textsc{Imitate} applies a subroutine \textsc{ListIdentifiableSpace} to list identifiable subspaces \(\Pi'\) w.r.t. \((\mathcal{G} \cup \{Y\}, \Pi, Y)\), following the observation made in Lem. 3. The implementation details of \textsc{ListIdentifiableSpace} are provided in \cite{15} Appendix C]. When an identifiable subspace \(\Pi'\) is found, \textsc{Imitate} tries to obtain a surrogate \(S\) w.r.t the diagram \(\mathcal{G}\) 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 \(\mathcal{G}\), a policy space \(\Pi\), an observational distribution \(P(o)\) and a subset \(S \subseteq O\). \(P(s)\) is p-imitating only if for any \(S' \subseteq S\), \(P(s)\) is p-imitating w.r.t. \((\mathcal{G}, \Pi, P(o))\).

We apply a subroutine \textsc{ListMinSep} in \cite{43} to enumerate minimal surrogates in \(O\) that d-separate \(\hat{X}\) and \(Y\) in the supergraph \(\mathcal{G} \cup \Pi'\). 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. \((\mathcal{G}, \Pi', S)\), i.e., \((\hat{S}, \Pi')\) 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 \({\pi : \mathcal{P}_X}\) w.r.t. \((\mathcal{G} \cup \{Y\}, \Pi, Y)\), which is associated with a minimal surrogate \({\hat{S}}\). Applying \textsc{Identify} confirms that \({\hat{S}}\) 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 \(\mathcal{M}_{(G,p)}\). 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 \(M \in \mathcal{M}_{(G,p)}\).

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

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 \(\mathcal{M}_{(G,p)}\) i.e., \(P(s|do(\pi))\) is identifiable w.r.t. \((\mathcal{G}, \Pi)\). We could thus express \(P(s|do(\pi); M)\) for any \(M \in \mathcal{M}_{(G,p)}\) 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{31, 33, 3}. As an
We demonstrate our algorithms on several synthetic datasets, including the above equation, with the actual distribution.

The interventional distribution

commonly exist. We refer readers to [15, Appendix D] for more experiments, details, and analysis.

imitate distributions over the expert’s reward in imitable (p-imitable) cases; and p-imitable instances

the actual reward distribution generated by an expert (bc) compared with the actual distribution $P(y)$ over the expert’s reward (opt).

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:

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

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 $P(\pi)$ could be computational challenging if some variables in $O$ are high-dimensional (e.g., $W$ in Fig. [1d]). Properties of the imitation instrument ($S, \Pi'$) suggest a practical approach to address this issue. Since $P(s|do(\pi))$ is identifiable w.r.t. ($\mathcal{G}, \Pi'$), by Def. [2] it remains invariant over the models in the hypothesis class $\mathcal{M}(\mathcal{G})$. This means that we could compute interventional probabilities $P(s|do(\pi); \hat{M})$ in an arbitrary model $\hat{M} \in \mathcal{M}(\mathcal{G})$; 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 $\mathcal{N}(\mathcal{G})$ be a parametrized subfamily of POSCMs in $\mathcal{M}(\mathcal{G})$. We could obtain an POSCM $\tilde{M} \in \mathcal{N}(\mathcal{G})$ such that its observational distribution $P(\pi; \tilde{M}) = P(\pi)$; an imitating policy $\pi$ is then computed in the parametrized model $\tilde{M}$. 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 $\mathcal{G}$, a policy space $\Pi$, and an observational distribution $P(\pi)$, let $(S, \Pi')$ be an instrument w.r.t. ($\mathcal{G}, \Pi$). If there exists a POSCM $\tilde{M} \in \mathcal{M}(\mathcal{G}, p)$ and a policy $\pi \in \Pi'$ such that $P(s|do(\pi); \tilde{M}) = P(s)$, then $P(y)$ is p-imitable w.r.t. ($\mathcal{G}, \Pi, P(\pi)$). Moreover, $\pi$ is an imitating policy for $P(y)$ w.r.t. ($\mathcal{G}, \Pi, P(\pi)$).

In practical experiments, we consider a parametrized family of POSCMs $\mathcal{N}(\mathcal{G})$ 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) [2, 27], we obtain a model $\hat{M} \in \mathcal{N}(\mathcal{G})$ satisfying the observational constraints $P(\pi; \hat{M}) = P(\pi)$. 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 on MNIST digits. In all experiments, we test our causal imitation method (ci); we apply Thm. [2] when there exists a $\pi$-backdoor admissible set; otherwise, Alg. [1] is used to leverage the observational distribution. As a baseline, we also include naive behavior cloning (bc) that mimics the observed conditional distribution $P(x|pa(\Pi))$, 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 a favorable characteristic of causal inference methods which finds an imitating policy, by exploiting quantitative knowledge contained in the observational values of velocity of the driving car. 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. 4b. 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. 4c 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 $P(x, w, s; M) = P(x, w, s)$. 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 $P(s|do(\pi); M)$ 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 highway 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 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
that they are inherently robust: for example, the definition of imitability Def. 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**


