Sound and Complete Incorporation of Local Causal Background Knowledge with Latent Variables

Anonymous Author(s)
Affiliation
Address
email

Abstract

Identifying causal relations is an important problem in various disciplines of science. When latent variables exist, ancestral graph is generally used to describe causal relations. Only a Markov equivalence class (MEC) is identifiable with observational data, represented by a partial ancestral graph (PAG) where the orientations of some edges are uncertain. Without further assumptions, we can only eliminate the uncertainty by incorporating background knowledge attainable through experiments or human experience. However, it is an open problem how to completely orient a PAG with background knowledge. The problem is fundamental due to its implication for the maximally identifiable causal relations based on the information. In this paper, we take a further step towards addressing it. We notice that the background knowledge in real tasks is usually in a *local* form. Given local background knowledge, we present *sound* and *complete* orientation rules, with which the PAG can be maximally oriented. As an application of the orientation rules, we propose the first general active learning framework for causal discovery in the presence of latent confounders, aiming to recover the true ancestral graph with as few interventions as possible. Specifically, we present a baseline maximal entropy criterion, equipped with Metropolis-Hastings sampling, to select the interventional target in each round. The experiments demonstrate the effectiveness and efficiency of the proposed framework to discover the ancestral graph.

1 Introduction

Causality has attracted tremendous attention in recent years, for its application on explainability [1], fairness [2, 3, 4], decision [5, 6, 7], and so on. In Pearl's causality framework [8], there are many studies on *causal discovery*, *i.e.*, learning the causal graph [9, 10, 11, 12, 13]. If we do not make some additional functional assumptions [14, 15, 16], or consider multiple or dynamic environments [17, 18, 19], generally, we can only learn a Markov equivalence class (MEC), which contains some uncertainty about the causal relations.

In light of the uncertainty in the graph, there are mainly two types of measures to reveal it. One is introducing experiments, through which we can collect additional interventional data to learn the graph [20, 21, 22, 23, 24, 25, 26]. The other is introducing human experience [27]. In fact, in many real tasks, it is unnecessary to learn the whole causal graph purely from data, as the human experience can usually tell us some causal relations. For example, a businessman is possibly interested in learning the causal relations between many variables involved in sales, in order to understand the causal relations or estimate causal effects accordingly to guide decisions. Here, causal relations such as price causes sales can be obtained directly according to the experience without any learning process. The causal knowledge acquired from the two types of measures above is collectively called *background knowledge*, or called *BK* for short.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

When BK is available in addition to a Markov equivalence class, a consequent theoretical problem is: how to completely orient the partially identified graph representing the MEC? It is a fundamental problem for its implication on the maximally oriented causal graph with the observational data and BK. In other words, it indicates the impossibility of some causal relations' identification given the data and BK. The difficulty of this problem results from the fact that, when we incorporate BK, in addition to the BK itself, we could also obtain some further causal relations. For example, edges can be oriented on the basis of some necessary conditions for the causal graph, such as there cannot be cyclic causal relations and the graph oriented by BK must belong to the MEC. It is hard to give a complete characterization for such additional causal knowledge. When we assume causal sufficiency, *i.e.*, there are no latent confounders that cause some observed variables simultaneously, Meek [28] established sound and complete rules, called *Meek rules*, for the orientation with BK. However, causal sufficiency is an untestable assumption in practice. And when we apply causality in subjects such as biology, sociology, and economics, it is quite often that there are some latent confounders. For example, the macroeconomic policy influences purchase price, the population of customers, and advertising cost, but it is hard for the businessman to evaluate it, thereby a latent confounder. In this case, it is a partial ancestral graph (PAG), which can be learned by FCI algorithm, that represents the MEC [29, 30, 31]. PAG can imply the *existence* of the causal relation between any two variables but not necessarily imply the causal *direction*. Unfortunately, how to completely incorporate BK into a PAG is still an open problem [31]. Andrews [32] show that FCI algorithm is complete with *tiered* BK, where all variables are partitioned into disjoint sets with explicit causal order. While in many cases, e.g., when BK is revealed by interventions, BK is not tiered.

In this paper, we take a further step towards addressing the problem. We consider the BK in a *local* form, which generally happens in real tasks no matter it is from experiments or human experiences. For example, when we make experiments and collect the data under intervention on X, for each variable V that has a causal relation with X, the interventional data can tell whether X causes V; and businessman often has enough domain knowledge about price, thus they usually know the causal relations between price and other variables, such as price causes sales and number of customers, and price is caused by stocks. We say the BK above is *local BK*, *i.e.*, if we have the BK regarding a variable X, we have the causal relations about X with any adjacent variable. We study the case when local BK is incorporated into a PAG, and present the *sound* and *complete* orientation rules.

The establishment of orientation rules compatible with local BK makes causal discovery by interventions possible when latent variables exist. We propose the first general active learning framework for causal discovery in the presence of latent variables, with the target of identifying the maximal ancestral graph (MAG). Considering that intervention is expensive in reality, we hope to achieve the target with as few interventions as possible. Hence we present a baseline maximal entropy criterion, equipped with Metropolis-Hastings sampling, to select the intervention variable such that we can learn more causal relations by each intervention. Our contributions in this paper are threefold:

(1) We present the sound and complete orientation rules to orient a PAG with local BK.

(2) We give the first active learning framework for causal discovery when latent variables exist.

(3) We introduce maximal entropy criterion equipped with Metropolis-Hastings sampling to select intervention variables as an instantiation of the proposed framework.

2 Preliminary

A graph $G = (\mathbf{V}, \mathbf{E})$ consists of a set of vertices $\mathbf{V} = \{V_1, \cdots, V_p\}$ and a set of edges $\mathbf{E}$. For any subset $\mathbf{V}' \subseteq \mathbf{V}$, the *subgraph induced by* $\mathbf{V}'$ is $G_{\mathbf{V}'} = (\mathbf{V}', \mathbf{E}_{\mathbf{V}'})$, where $\mathbf{E}_{\mathbf{V}'}$ is the set of edges in $\mathbf{E}$ whose both endpoints are in $\mathbf{V}'$. For a graph G, $\mathbf{V}(G)$ denotes the set of vertices in G. G is a complete graph if there is an edge between any two vertices. The subgraph induced by an empty set is trivially a complete graph. $G[-\mathbf{V}']$ denotes the subgraph $G_{\mathbf{V}\setminus\mathbf{V}'}$ induced by $\mathbf{V}\setminus\mathbf{V}'$. Usually, bold letter (*e.g.*, $\mathbf{V}$) denotes a set of vertices and normal letter (*e.g.*, V) denotes a vertex. A graph is *chordal* if any cycle of length four or more has a chord, which is an edge joining two vertices that are not adjacent in the cycle. If $G = (V, E)$ is chordal, the subgraph of G induced by $V' \subseteq V$ is chordal.

A graph G is *mixed* if the edges in G are either directed $\rightarrow$ or bi-directed $\leftrightarrow$. The two ends of an edge are called *marks* and have two types *arrowhead* or *tail*. A graph is a *partial mixed graph* (*PMG*) if it contains directed edges, bi-directed edges, and edges with *circles* ($\circ$). The circle implies that the mark here could be either arrowhead or tail but is indefinite. V_i is *adjacent* to V_j in G if there is an

edge between V_i and V_j. A path in a graph G is a sequence of distinct vertices $\langle V_0, \cdots, V_n \rangle$ such that for $0 \leq i \leq n-1$, V_i and V_{i+1} are adjacent in G. An edge in the form of $V_i \circ\!\!-\!\!\circ V_i$ is a *circle edge*. The circle component in G is the subgraph consisting of all the $\circ\!\!-\!\!\circ$ edges in G. Denote the set of vertices adjacent to V_i in G by $\text{Adj}(V_i, G)$. A vertex V_i is a *parent* of a vertex V_j if there is $V_i \rightarrow V_j$. A *directed path* from V_i to V_j is a path comprised of directed edges pointing to the direction of V_j. A *possible directed path* from V_i to V_j is a path without arrowhead at the mark near V_i on each edge. V_i is an *ancestor/possible ancestor* of V_j if there is a *directed path/possible directed path* from V_i to V_j or $V_i = V_j$. V_i is a *descendant/possible descendant* of V_j if there is a *directed path/possible directed path* from V_j to V_i or $V_j = V_i$. Denote the set of *parent/ancestor/possible ancestor/descendant/possible descendant* of V_i in G by $\text{Pa}(V_i, G)$/$\text{Anc}(V_i, G)$/$\text{PossAn}(V_i, G)$/$\text{De}(V_i, G)$/$\text{PossDe}(V_i, G)$. If $V_i \in \text{Anc}(V_j, G)$ and $V_i \leftarrow V_j$/$V_i \leftrightarrow V_j$, it forms a *directed cycle/almost directed cycle*. $*$ is a wildcard that denotes any of the marks (arrowhead, tail, and circle). We make a convention that when an edge is in the form of $\circ\!\!-\!\!*$, the $*$ here cannot be a tail since in this case the circle can be replaced by an arrowhead due to the assumption of no selection bias.

A non-endpoint vertex V_i is a collider on a path if the path contains $*\!\!\rightarrow V_i \leftarrow\!\!*$. A path p from V_i to V_j is a *collider path* if V_i and V_j are adjacent or all the passing vertices are colliders on p. p is a *minimal path* if there are no edges between any two non-consecutive vertices. A path p from V_i to V_j is a *minimal collider path* if p is a collider path and there is not a proper subset $\mathbf{V}'$ of the vertices in p such that there is a collider path from V_i to V_j comprised of $\mathbf{V}'$. A triple $\langle V_i, V_j, V_k \rangle$ on a path is unshielded if V_i and V_k are not adjacent. p is an *uncovered path* if every consecutive triple on p is unshielded. A path p is a *minimal possible directed path* if p is minimal and possible directed.

A mixed graph is an *ancestral graph* if there is no directed or almost directed cycle (since we assume no selection bias, we do not consider undirected edges in this paper). An ancestral graph is a *maximal ancestral graph (MAG, denoted by $\mathcal{M}$)* if it is *maximal*, i.e., for any two non adjacent vertices, there is a set of vertices that *m-separates* them [29]. A path p between X and Y in an ancestral graph G is an *inducing path* if every non-endpoint vertex on p is a collider and meanwhile an ancestor of either X or Y. An ancestral graph is maximal if and only if there is no inducing path between any two non-adjacent vertices [29].

In an MAG, a path $p = \langle X, \cdots, W, V, Y \rangle$ is a *discriminating path for V* if (1) X and Y are not adjacent, and (2) every vertex between X and V on the path is a collider on p and a parent of Y. Two MAGs are *Markov equivalent* if they share the same *m-separations*. A class comprised of all Markov equivalent MAGs is a *Markov equivalence class (MEC)*. We use a *partial ancestral graph (PAG, denoted by $\mathcal{P}$)* to denote an MEC, where a tail/arrowhead occurs if the corresponding mark is tail/arrowhead for all Markov equivalent MAGs, and a circle occurs otherwise.

For a PMG $\mathbb{M}$ that is obtained from a PAG $\mathcal{P}$ by orienting some circles to either arrowheads or tails, an MAG $\mathcal{M}$ is *consistent to the PMG* $\mathbb{M}$ if (1) the non-circle marks in $\mathbb{M}$ are also in $\mathcal{M}$, and (2) $\mathcal{M}$ is in the MEC represented by $\mathcal{P}$. We say $\mathcal{M}$ *is consistent to the BK* if $\mathcal{M}$ is with the orientations dictated by the BK.

3 The rules to orient a PAG with local background knowledge

In this section, we present the sound and complete orientation rules to orient a PAG $\mathcal{P}$ with local background knowledge (BK), where $\mathcal{P}$ is learned by observational data [9, 31] and $\mathbf{V}(\mathcal{P}) = \{V_1, V_2, \cdots, V_d\}$. The *local BK* regarding X means that the BK directly implies and only directly implies all the true marks at X, denoted by $BK(X)$. We assume the absence of selection bias and that the BK is correct. The correctness indicates that there exists an MAG consistent to $\mathcal{P}$ and the BK. Without loss of generality, we suppose the local BK is regarding $V_1, V_2, \cdots, V_k, 1 \leq k \leq d$. That is, for any vertex $X \in V_1, V_2, \cdots, V_k$, all the marks at X are known according to the local BK; and for any vertex $X \in V_{k+1}, \cdots, V_d$, the local BK does not directly imply any marks at X.

First, we show the orientation rules to incorporate local BK. They follow the rules of Zhang [31] for learning a PAG but with one replacement and one addition. Due to the page limit, we do not list them here but the replaced and additional ones. See Appendix A for the rules proposed by Zhang [31].

$\mathcal{R}_4'$: If $\langle K, \cdots, A, B, R \rangle$ is a discriminating path between K and R for B, and $B \circ\!\!-\!\!* R$, then orient $B \circ\!\!-\!\!* R$ as $B \rightarrow R$.

$\mathcal{R}_{11}$: If $A \rightarrow\!\!* B$, then $A \rightarrow B$.

Algorithm 1: Update a PMG with local background knowledge

Input: A PMG $\mathbb{M}_i$, $BK(X)$

Output: Updated graph $\mathbb{M}_{i+1}$

1 For any $K \in \text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}])$ and any $T \in \mathbf{C}$ such that $K \circ\!\!-\!\!* T$ in $\mathbb{M}_i$, orient $K \leftarrow\!\!* T$ (the mark at T remains); for all $K \in \text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}])$ such that $X \circ\!\!-\!\!* K$, orient $X \rightarrow K$;

2 Orient the subgraph $\mathbb{M}_i[\text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}])\backslash\{X\}]$ as follows until no feasible updates: for any two vertices V_l and V_j such that $V_l \circ\!\!-\!\!\circ V_j$, orient it as $V_l \rightarrow V_j$ if **(i)** $\mathcal{F}_{V_l}\backslash\mathcal{F}_{V_j} \neq \emptyset$ or **(ii)** $\mathcal{F}_{V_l} = \mathcal{F}_{V_j}$ as well as there is a vertex $V_m \in \text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}])\backslash\{X\}$ not adjacent to V_j such that $V_m \rightarrow V_l \circ\!\!-\!\!\circ V_j$;

3 Apply the orientation rules until the graph is closed under the orientation rules.

Prop. 1 implies the soundness of $\mathcal{R}_4'$ to orient a PAG $\mathcal{P}$ or a PMG obtained from $\mathcal{P}$ with local BK. See Appendix A for the proof. Roughly speaking, if there is $B \leftarrow\!\!* R$ in an MAG consistent with the case of $\mathcal{R}_4'$, there must be a minimal collider path between K and R across K, in which case $B \leftarrow\!\!* R$ *should have been* identified in the PAG according to Zhao et al. [33], Zhang [31], contradiction. $\mathcal{R}_{11}$ is immediate due to no selection bias assumption. In the following, we make a convention that when we say the orientation rules, they refer to $\mathcal{R}_1 - \mathcal{R}_3, \mathcal{R}_8 - \mathcal{R}_{10}$ of Zhang [31] and $\mathcal{R}_4', \mathcal{R}_{11}$. A PMG is *closed under the orientation rules* if the PMG cannot be oriented further by the orientation rules.

Proposition 1. *Given a PAG $\mathcal{P}$, for any PMG $\mathbb{M}$ that is obtained from $\mathcal{P}$ by orienting some circles in $\mathcal{P}$ (or $\mathbb{M} = \mathcal{P}$), $\mathcal{R}_4'$ is sound to orient $\mathbb{M}$ with local background knowledge.*

Next, we will prove the completeness of the proposed orientation rules. It is somewhat complicated. We first give a roadmap for the proof process. There are mainly two parts. The first is that we present an algorithm to orient $\mathcal{P}$ with the local BK regarding $V_1, V_2, \cdots, V_k$. The second part is to prove that the algorithm can be achieved by the proposed orientation rules. Combining these two parts, we conclude the orientation rules are complete to orient a PAG. The construction of the algorithm and the corresponding proof for the completeness of the algorithm in the first step is the most difficult part. To achieve the construction, we divide the whole process of orienting a PAG with BK regarding $V_1, V_2, \cdots, V_k$ into k steps. Beginning from the PAG $\mathcal{P}$ ($\mathcal{P}$ is also denoted by $\mathbb{M}_0$), in the $(i+1)$-th $(0 \leq i \leq k-1)$ step we obtain a PMG $\mathbb{M}_{i+1}$ from $\mathbb{M}_i$ by incorporating $BK(V_{i+1})$ and orienting some other circles further. To obtain the updated graph in each step, we propose an algorithm orienting a PMG with local BK incorporated in this step. Repeat this process by incorporating $BK(V_1), BK(V_2), \ldots, BK(V_k)$ sequentially, we obtain the PMG with incorporated BK regarding $V_1, \cdots, V_k$. We will prove that the k-step algorithm to orient PAG with local BK regarding $V_1, \cdots, V_k$ is complete, by an induction step that if the first i-step algorithm is complete to update the PAG $\mathcal{P}$ with BK regarding $V_1, \cdots, V_i$, then the $(i+1)$-step algorithm is complete to update $\mathcal{P}$ with BK regarding $V_1, \cdots, V_{i+1}$. Hence the proof in the first part completes. In the second part, we show that the k-step algorithm can be achieved by the proposed orientation rules. Since the rules are also proved sound, they are sound and complete to orient the PAG with local BK.

We present Alg. 1 to obtain $\mathbb{M}_{i+1}$ from $\mathbb{M}_i$ by incorporating $BK(V_{i+1})$. For brevity, we denote V_{i+1} by X, and introduce a set of vertices $\mathbf{C}$ defined as $\mathbf{C} = \{V \in \mathbf{V}(\mathcal{P}) \mid V *\!\!\rightarrow X \in BK(X)\}$ to denote the vertices whose edges with X will be oriented to ones with arrowheads at X according to $BK(X)$. In $\mathbb{M}_{i+1}$, there is $X \leftarrow\!\!* V$ for $V \in \mathbf{C}$ and $X \rightarrow\!\!* V$ for $V \in \{V \in \mathbf{V}(\mathcal{P}) \mid V *\!\!-\!\!\circ X \text{ in } \mathbb{M}_i\}\backslash\mathbf{C}$ oriented directly by $BK(X)$. We define $\mathcal{F}_{V_l}^{\mathbb{M}_i} = \{V \in \mathbf{C} \cup \{X\} \mid V *\!\!-\!\!\circ V_l \text{ in } \mathbb{M}_i\}$ for any $V_l \in \text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}])\backslash\{X\}$, which is denoted by $\mathcal{F}_{V_l}$ for short. $\mathcal{F}_{V_l}$ denotes the vertices in $\mathbf{C} \cup \{X\}$ whose edges with V_l are oriented to ones with arrowheads at V_l in the first step of Alg. 1.

In the first step of Alg. 1, the orientation at X follows $BK(X)$, and the orientation at the vertices apart of X is motivated as the necessary condition for the ancestral property. If there is an oriented edge $K \rightarrow T$ in the case of the first step, then no matter how we orient the other circles, there will be a directed or almost directed cycle, unless we introduce new unshielded colliders (which takes new conditional independences relative to those in $\mathcal{P}$), both of which are evidently invalid to obtain an MAG in the MEC represented by $\mathcal{P}$. And the orientation in the second step is motivated as the necessary condition for that there are no new unshielded colliders in the oriented graph relative to the PAG $\mathcal{P}$. If there is an MAG where there is an inconsistent edge with the edge oriented in this

step, then there must be new unshielded colliders relative to $\mathcal{P}$, which implies that the MAG is not consistent to $\mathcal{P}$. The third step orients some other circles based on the updated structure.

Then, we present the key induction result in Thm. 1 for the graph obtained by Alg. 1 in each step. Due to the page limit, we only show a proof sketch, with a detailed version in Appendix B.1. Then with Thm. 1, we directly conclude that k-step algorithm is complete to orient the PAG with the local BK regarding $V_1, \ldots, V_k$ in Cor. 1.

Theorem 1. Given i, suppose $\mathbb{M}_s, \forall s \in \{0, 1, \ldots, i\}$ satisfies the five following properties:

(Closed) $\mathbb{M}_s$ is closed under the orientation rules.

(Invariant) The arrowheads and tails in $\mathbb{M}_s$ are invariant in all the MAGs consistent to $\mathcal{P}$ and BK regarding $V_1, \ldots, V_s$.

(Chordal) The circle component in $\mathbb{M}_s$ is chordal.

(Balanced) For any three vertices A, B, C in $\mathbb{M}_s$, if $A{*\!\!\rightarrow} B \circ\!\!{*} C$, then there is an edge between A and C with an arrowhead at C, namely, $A{*\!\!\rightarrow} C$. Furthermore, if the edge between A and B is $A \rightarrow B$, then the edge between A and C is either $A \rightarrow C$ or $A{\circ\!\!\rightarrow} C$ (i.e., it is not $A \leftrightarrow C$).

(Complete) For each circle at vertex A on any edge $A \circ\!\!{*} B$ in $\mathbb{M}_s$, there exist MAGs $\mathcal{M}_1$ and $\mathcal{M}_2$ consistent to $\mathcal{P}$ and BK regarding $V_1, \ldots, V_s$ with $A \leftarrow\!\!{*} B \in \mathbf{E}(\mathcal{M}_1)$ and $A \rightarrow B \in \mathbf{E}(\mathcal{M}_2)$.

Then the PMG $\mathbb{M}_{i+1}$ obtained from $\mathbb{M}_i$ with $BK(V_{i+1})$ by Alg. 1 also satisfies the five properties.

Proof sketch: For brevity, we denote V_{i+1} by X. **(A)** The closed property holds due to the third step of Alg. 1.**(B)** The invariant property holds because all the orientations in Alg. 1 either follow $BK(X)$ or are motivated as the necessary condition for the ancestral property and the fact that there cannot be new unshielded colliders introduced relative to $\mathbb{M}_i$. **(C)** The chordal property is proved based on the fact that only the first two steps of Alg. 1 possibly introduce new arrowheads, while the third step will only orient circles to tails, which is proved in Lemma 12 in Appendix B.1. With this fact, it suffices to prove that the circle component in the graph obtained after the first two steps is chordal. Denote the graph after the first two steps by $\bar{\mathbb{M}}_{i+1}$. We can prove that the circle components in $\bar{\mathbb{M}}_{i+1}[\text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}])]$ and in $\bar{\mathbb{M}}_{i+1}[-\text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}])]$ are chordal, respectively. Since there are no circle edges connecting $\text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}])$ and $\mathbf{V} \backslash \text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}])$ (otherwise it has been oriented in the first step of Alg. 1), we conclude the desired result. **(D)** The balanced property of $\mathbb{M}_{i+1}$ is proved based on three facts that (1) in Alg. 1, if we transform a circle to arrowhead at V, then $V \in \text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}])$; (2) if there is $A \in \text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}])$ and $A \circ\!\!{*} B$, $B \notin \mathbf{C}$, in $\mathbb{M}_{i+1}$, then $B \in \text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}])$; (3) $\mathbb{M}_i$ satisfies the balanced property. We can prove that it is impossible that there is a sub-structure $V_i{*\!\!\rightarrow} V_j \circ\!\!{*} V_k$ where V_i is not adjacent to V_k or there is $V_i {*\!\!-\!\!\circ} V_k$ in $\mathbb{M}_{i+1}$ by discussing whether V_i, V_j, V_k belongs to $\text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}])$. **(E)** The completeness property is proved by showing two results: (1) in the MAGs consistent to $\mathcal{P}$ and local BK regarding $V_1, \cdots, V_{i+1}$, each circle edge $A \circ\!\!-\!\!\circ B$ can be oriented as both $A \rightarrow B$ and $A \leftarrow B$, and each edge $A{\circ\!\!\rightarrow} B$ can be oriented as $A \rightarrow B$; (2) each edge $A{\circ\!\!\rightarrow} B$ can be oriented as $A \leftrightarrow B$ in an MAG consistent to $\mathcal{P}$ and local BK regarding $V_1, \cdots, V_{i+1}$. In this part, the most difficult part is to prove the first result, with which the second result can be proved directly following the proof process of Thm. 3 of Zhang [31]. In the proof for the first result, we show that any MAG obtained from $\mathbb{M}_{i+1}$ by transforming the edges as $A{\circ\!\!\rightarrow} B$ to $A \rightarrow B$ and the circle component into a DAG without new unshielded colliders is consistent to $\mathcal{P}$ and local BK regarding $V_1, \ldots, V_{i+1}$. If not, we can always find an MAG obtained from $\mathbb{M}_i$ by transforming the edges as $A{\circ\!\!\rightarrow} B$ to $A \rightarrow B$ and the circle component into a DAG without new unshielded colliders that is not consistent to $\mathcal{P}$ and local BK regarding $V_1, \ldots, V_i$. By induction, there is an MAG obtained from $\mathcal{P}$ by transforming the edges as $A{\circ\!\!\rightarrow} B$ to $A \rightarrow B$ and the circle component into a DAG without new unshielded colliders that is not consistent to $\mathcal{P}$, contradiction with Thm. 2 of Zhang [31]. We conclude the first result. □

Corollary 1. The k-step algorithm from $\mathbb{M}_0 (= \mathcal{P})$ to $\mathbb{M}_k$ is sound and complete. That is, the non-circle marks in $\mathbb{M}_k$ are invariant in all the MAGs consistent to $\mathcal{P}$ and BK regarding $V_1, \ldots, V_k$. And for each circle in $\mathbb{M}_k$, there exist both MAGs with an arrowhead and MAGs with a tail here that are consistent to $\mathcal{P}$ and BK regarding $V_1, \ldots, V_k$.

Proof. Previous studies [30, 31] show that the last four properties in Thm. 1 are fulfilled in PAG, the case in $\mathcal{R}_4'$ will never happen in $\mathcal{P}$ because such circles have been oriented by $\mathcal{R}_4$ in the process of learning $\mathcal{P}$, and the case in $\mathcal{R}_{11}$ is never triggered by the rules to learn $\mathcal{P}$. Hence $\mathcal{P}$ satisfies the five

properties. With the induction step implied by Thm 1, we directly conclude that $\mathbb{M}_k$ satisfies the five properties, thereby satisfying the invariant and complete property. $\square$

Theorem 2. The orientation rules are sound and complete to orient a PAG with the local background knowledge regarding $V_1, \ldots, V_k$.

Proof. The soundness of $\mathcal{R}_4'$ is shown by Prop. 1. The soundness of other rules immediately follows Thm. 4.1 of Ali et al. [30] and Thm. 1 of Zhang [31]. We do not show the details. Roughly speaking, the violation of these rules will lead to that there are new unshielded colliders or directed or almost directed cycles in the oriented graph relative to $\mathcal{P}$. The main part is to prove the completeness.

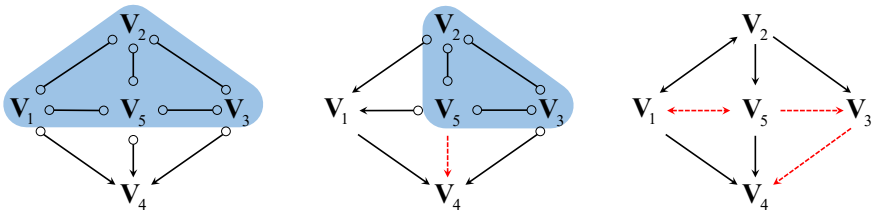
According to Cor. 1, it suffices to prove that in each step by Alg. 1 to incorporate $BK(X)$ into a PMG $\mathbb{M}_i$, the orientations in Alg. 1 either follow $BK(X)$ directly, or can be achieved by the proposed orientation rules. The orientation in the second step of Alg. 1 can be achieved by $\mathcal{R}_1$, because no matter $\mathcal{F}_{V_l} \backslash \mathcal{F}_{V_j} \neq \emptyset$ or $V_m \to V_l \circ\!\!-\!\!\circ V_j$, there is $F \in \mathcal{F}_{V_l} \backslash \mathcal{F}_{V_j}$ or $F = V_m$ respectively such that $F *\!\!\to V_l \circ\!\!-\!\!\circ V_j$ where F is not adjacent to V_j. The orientation in the third step naturally follows the orientation rules. For the orientation in the first step, $X \leftrightarrow\!\!* V$ for $V \in \mathbf{C}$ is dictated by $BK(X)$, and $X \to V$ for $V \in \{V \in \mathbf{V}(\mathcal{P}) \mid X \circ\!\!-\!\!* V\} \backslash \mathbf{C}$ is obtain from $X -\!\!* V$ dictated by $BK(X)$ and $\mathcal{R}_{11}$. The remaining part is to prove for $K \in \text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}]) \backslash \{X\}$ and $T \in \mathbf{C}$, if there is $K \circ\!\!-\!\!* T$ in $\mathbb{M}_i$, $K \leftrightarrow\!\!* T$ can be oriented by the proposed orientation rules when we incorporate $BK(X)$.

Due to $K \in \text{PossDe}(X, \mathbb{M}_i[-\mathbf{C}]) \backslash \{X\}$, there is a possible directed path from X to K that does not go through $\mathbf{C}$. According to Lemma 2 in Appendix B.1, there is a minimal possible directed path $p = \langle X(= F_0), F_1, \ldots, K(= F_t) \rangle, t \geq 1$ where each vertex does not belong to $\mathbf{C}$. Hence $X \to F_1$ is oriented by $BK(X)$ and $\mathcal{R}_{11}$ unless $X \to F_1$ has been in $\mathbb{M}_i$. Hence, $X \to F_1 \to \cdots \to F_t$ can be oriented by $\mathcal{R}_1$ after incorporating $BK(X)$ unless they have been in $\mathbb{M}_i$. If $t = 1$, there is $T *\!\!\to X \to K$, thus $K \leftrightarrow\!\!* T$ can be oriented by $\mathcal{R}_2$. Next, we consider the case when $t \geq 2$.

We first prove that for any $F_m \in F_1, \ldots, F_t, t \geq 2$, F_m is adjacent to T, and there is not $F_m \to T$ in $\mathbb{M}_i$. Suppose F_m is not adjacent to T, there must be a sub-structure of $\mathbb{M}_i$ induced by $F_{m-s}, F_{m-s+1}, \ldots, F_{m+l}, T, 1 \leq s \leq m, 1 \leq l \leq t-m$, such that T is only adjacent to F_{m-s} and F_{m+l} in this sub-structure. There are at least four vertices in this sub-structure. Hence there must be an unshielded collider (denoted by UC for short) in this sub-structure in $\mathcal{P}$, otherwise no matter how we orient the circle there is either a new UC relative to $\mathcal{P}$ or a directed or almost directed cycle there. Since p is possibly directed, the UC is at either F_{m+l} or T (*i.e.*, $*\!\!\to F_{m+l}($ or $T) \leftrightarrow\!\!*$). If there is an UC at F_{m+l}, $T *\!\!\to F_{m+l}$ and $F_{m+l-1} *\!\!\to F_{m+l}$ are identified in $\mathcal{P}$. Thus $F_{m+l} \to F_{m+l+1} \cdots \to F_t$ is identified in $\mathcal{P}$. Due to the completeness of FCI algorithm to learn $\mathcal{P}$, there is $K \leftrightarrow\!\!* T$ in $\mathcal{P}$, because there is not an MAG with $K \to T$ (there has been $T *\!\!\to F_{m+l} \to \cdots \to K$ in $\mathcal{P}$). Hence there is $K \leftrightarrow\!\!* T$ in $\mathbb{M}_i$, contradicting with $K \circ\!\!-\!\!* T$ in $\mathbb{M}_i$. If there is not a UC at F_{m+l}, UC can only be at T. Thus $F_{m-s} *\!\!\to T \leftrightarrow\!\!* F_{m+l}$ is identified in $\mathcal{P}$. Since p is possibly directed, F_{m+l-1} is not adjacent to T, and there is not a UC at F_{m+l} in the sub-structure, there cannot be $F_{m+l} \leftrightarrow T$ in $\mathcal{P}$. Hence the path $\langle F_{m-s}, F_{m-s+1}, \ldots, F_{m+l}, T \rangle$ in $\mathcal{P}$ is an uncovered possible directed path, $F_{m-s} \to T$ is identified in $\mathcal{P}$ (otherwise $\mathcal{R}_9$ applies). When incorporating $BK(X)$, there is a (almost) directed cycle $T *\!\!\to X \to \cdots \to F_{m-s} \to T$, contradicting with the correctness of BK. Hence, F_m is adjacent to T. Similarly, if $F_m \to T$ in $\mathbb{M}_i$, there is $T *\!\!\to X \to \cdots \to F_m \to T$, impossibility.

Finally, since F_1 is adjacent to T, and $T *\!\!\to X \to F_1$ is oriented according to $BK(X)$, there is $T *\!\!\to F_1$ oriented by $\mathcal{R}_2$ unless $T *\!\!\to F_1$ has been in $\mathbb{M}_i$. Hence there is always $T *\!\!\to F_1$ by the orientation rules. Consider $T *\!\!\to F_1 \to F_2$, there is $T *\!\!\to F_2$ oriented by $\mathcal{R}_2$ unless $T *\!\!\to F_2$ has been in $\mathbb{M}_i$. Repeat the process for $F_3, F_4, \ldots, F_t(= K)$, we can prove that if there is $F_t(= K) \circ\!\!-\!\!* T$ in $\mathbb{M}_i$, there is $T *\!\!\to F_t(= K)$ oriented by $\mathcal{R}_2$. Alg. 1 is thus achievable by the orientation rules. $\square$

We give an example in Fig. 1. Suppose we obtain a PAG as Fig. 1(a) with observational data and have the local BK regarding V_1, V_2. We divide the whole process of obtaining a PMG from $\mathcal{P}$ with the local BK into obtaining $\mathbb{M}_1$ from $\mathcal{P}$ with $BK(V_1)$ by Alg. 1 and then obtaining $\mathbb{M}_2$ from $\mathbb{M}_1$ with $BK(V_2)$ by Alg. 1. $\mathbb{M}_1$ and $\mathbb{M}_2$ are shown in Fig. 1(b) and 1(c), respectively. It is not hard to verify that all of $\mathcal{P}$, $\mathbb{M}_1$, $\mathbb{M}_2$ satisfy the closed, chordal, and balanced properties defined in Thm 1. Note if we do not consider $\mathcal{R}_4'$, the edge colored red in Fig. 1(b) cannot be oriented. Fig. 1(a) also shows a case where $BK(V_1)$ is not tiered [32]. The reason is that the vertices V_1, V_4, V_5 cannot partitioned into disjoint subsets with explicit causal order because V_1 and V_4 belong to different subsets according to $BK(V_1)$ but V_5 has ancestor relation with neither V_1 nor V_4.



(a) PAG $\mathcal{P}$(=$\mathbb{M}_0$) (b) $\mathbb{M}_1$ (c) $\mathbb{M}_2$

Figure 1: Fig. 1(a) depicts a PAG $\mathcal{P}$, with the local BK regarding V_1 in the form of $V_1 \leftrightarrow\!* V_2, V_1 \leftarrow\!* V_5, V_1 \rightarrow V_4$ and the local BK regarding V_2 in the form of $V_2 \leftrightarrow\!* V_1, V_2 \rightarrow V_3, V_2 \rightarrow V_5$. Circle components are indicated by shading. The edge oriented by the rules are denoted by red dashed line.

4 Active learning framework for causal discovery with latent variables

The establishment of the orientation rules for orienting a PAG with local BK makes causal discovery by interventions possible in the presence of latent variables. Hence, based on the theoretical results, we propose an active learning framework for causal discovery in the presence of latent confounders, with the target of learning the MAG with as fewer interventions as possible. The framework is comprised of three stages. In Stage 1, we learn a PAG with observational data. In Stage 2, we select a singleton variable $X \in V_1, \ldots, V_d$ to intervene and collect the interventional data. In Stage 3, we learn causal relations with the data. For each edge $X \circ\!\!-\!\!* V_i$, the circle at X can be learned by a two-sample test on whether the interventional distribution of V_i equals to the observational one. There is $X \leftrightarrow\!* V_i$ learned if they are equal, and $X -\!\!* V_i$ otherwise. Hence, the knowledge taken by the interventional data is local. We repeat the second and third stages until we identify the MAG. Since the orientation rules are complete, the graph can be updated completely by each intervention. The only remaining problem is how to select the intervention variable in Stage 2.

Considering that the whole process is sequential, we only focus on the intervention variable selection in one round. Without loss of generality, suppose we have obtained a PMG $\mathbb{M}_i$ by i interventions on $V_1, V_2, \ldots, V_i$ where some circles exist, and will select a variable from $\{V_{i+1}, \ldots, V_d\}$ to intervene. We adopt the maximum entropy criterion [20]. For $\mathbb{M}_i$, we select the variable X that maximizes

$$H_X = -\sum_{j=1}^{M} \frac{l_j}{L} \log \frac{l_j}{L}, \tag{1}$$

where j is an index for a local structure of X (a local structure of X denotes a definite orientation of the marks at X), M denotes the number of different local structures, l_j denotes the number of MAGs consistent to $\mathbb{M}_i$ which has the j-th local structure of X, and L denotes the total number of MAGs consistent to $\mathbb{M}_i$. Intuitively, the maximum entropy criterion is devoted to selecting the intervention variable X such that there is a similar number of MAGs with each local structure of X and as more as possible local structures of X. A justification for intervening on such a variable is that we hope to have s small space of MAGs after the intervention no matter what the true local structure of X is.

However, it is hard to count the number of MAGs consistent to $\mathbb{M}_i$ with a definite local structure. Even in causal sufficiency setting, implementing such operation (generally called counting maximally oriented partial DAGs) is #P-complete [34]. Considering DAG is a special case for MAG, the counting of MAGs is harder. Hence, we adopt a sampling method based on Metropolis-Hastings (MH) algorithm [35], to uniformly sample from the space of MAGs. The algorithm begins from an MAG consistent to $\mathbb{M}_i$, and in each round we transform the MAG to a candidate MAG and decide to accept or reject it with some probability. If any two states can be transformed to each other in limited steps, a stationary distribution equal to the desired distribution can be obtained by MH algorithm [36]. Since any MAG can be transformed to another Markov equivalent MAG [37], MH algorithm is valid to sample MAGs uniformly from the space of MAGs consistent to $\mathbb{M}_i$. Here, we introduce an important result of Zhang and Spirtes [37] for MAGs transformation in Prop. 2.

Proposition 2 (Zhang and Spirtes [37]). *Let $\mathcal{M}$ be an arbitrary MAG, and $A \rightarrow B$ an arbitrary directed edge in $\mathcal{M}$. Let $\mathcal{M}'$ be the graph identical to $\mathcal{M}$ except that the edge between A and B is $A \leftrightarrow B$. $\mathcal{M}'$ is an MAG Markov equivalent to $\mathcal{M}$ if and only if*

(1) there is no directed path from A to B other than $A \to B$ in $\mathcal{M}$;

(2) for any $C \to A$ in $\mathcal{M}$, $C \to B$ is also in $\mathcal{M}$; and for any $D \leftrightarrow A$ in $\mathcal{M}$, either $D \to B$ or $D \leftrightarrow B$ is in $\mathcal{M}$;

(3) there is no discriminating path from A on which B is the endpoint adjacent to A in $\mathcal{M}$.

Given an MAG $\mathcal{M}$, let $S(\mathcal{M})$ denote the set of MAGs that can be obtained from $\mathcal{M}$ by transforming one bi-directed edge to directed edge or one directed edge to bi-directed edge according to Prop. 2. Denote the cardinality of $S(\mathcal{M})$ by $|S(\mathcal{M})|$. We set the probability $Q(\mathcal{M}' \mid \mathcal{M})$ of an MAG $\mathcal{M}$ transformed to another MAG $\mathcal{M}' \in S(\mathcal{M})$ as $1/|S(\mathcal{M})|$. Hence, the acceptance ratio ρ that is used to decide whether to accept or reject the candidate is

$$\rho = \min\left(1, \frac{p(\mathcal{M}')Q(\mathcal{M} \mid \mathcal{M}')}{p(\mathcal{M})Q(\mathcal{M}' \mid \mathcal{M})}\right) = \min\left(1, \frac{|S(\mathcal{M})|}{|S(\mathcal{M}')|}\right).$$

We propose Alg. 2 to select the intervention variable X. As shown by Lemma 15.1 in Appendix B.1, the graph $\mathcal{M}_0$ is an MAG consistent to $\mathbb{M}_i$. From Line 2-Line 6, we execute MH algorithm to sample L MAGs. Then, we estimate the entropy by (1) and select X from Line 8-Line 13.

Algorithm 2: Intervention variable selection based on maximum entropy criterion with MH alg.

Input: A PMG $\mathbb{M}_i$ oriented based on $\mathcal{P}$ and BK regarding $V_1, \ldots, V_i$, number of MAGs L
Output: The selected intervention variable X

1 Obtain an MAG $\mathcal{M}_0$ based on $\mathbb{M}_i$ by transforming $\circ\!\!\to$ to $\to$ and the circle component into a DAG without new unshielded colliders;
2 **for** $t = 1, 2, \ldots, L$ **do**
3 Sample an MAG $\mathcal{M}'$ from $S(\mathcal{M}_{t-1})$;
4 $\rho = \min(1, \frac{|S(\mathcal{M}_{t-1})|}{|S(\mathcal{M}')|})$;
5 Sample u from uniform distribution $U[0, 1]$;
6 **if** $u \leq \rho$ **then** $\mathcal{M}_t = \mathcal{M}'$ **else** $\mathcal{M}_t = \mathcal{M}_{t-1}$;
7 $s \leftarrow 0$, $X \leftarrow \emptyset$;
8 **for** $V_j = V_{k+1}, \ldots, V_d$ **do**
9 Denote $\mathbf{V}(V_j) = \{V \in \mathbf{V}(\mathbb{M}_i) \mid V_j \circ\!\!\!*\, V \text{ in } \mathbb{M}_i\}$;
10 For each possible local structure $\mathcal{L}_k$ of V_j, $1 \leq k \leq 2^{|\mathbf{V}(V_j)|}$, we count the number $\mathcal{N}_k$ of the appearance of $\mathcal{L}_k$ in L MAGs;
11 $s' = -\sum_{k=1}^{2^{|\mathbf{V}(V_j)|}} \frac{\mathcal{N}_k}{L} \log \frac{\mathcal{N}_k}{L}$;
12 **if** $s \leq s'$ **then** $X \leftarrow V_j$, $s \leftarrow s'$;
13 **return** X.

5 Experiments

In this section, we evaluate the performance of the three-stage active learning framework. We generate 100 Erdös-Rényi random DAGs for each setting, where the number of nodes $d \in \{6, 8, 10, 12, 14, 16\}$ and the probability of including each edge $p \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$. The weight of each edge is drawn from $\mathcal{U}[1, 2]$. We generate 10000 samples from the corresponding linear structural equations and take three nodes as latent variables and the others as observed ones. In the implementation of the MH algorithm, we discard the first 500 sampled MAGs and collect the following 1000 MAGs. For each intervention variable X, we collect 10000 samples under $do(X = 2)$, and learn the circles at X by two-sample tests with a significance level of 0.05. We present further details of the experiments, evaluation metrics, and complete experimental results in Appendix C due to limited space.

We first evaluate the three stages respectively. We compare the proposed MCMC strategy with a baseline random strategy where we randomly select one node with circles in each round. In Stage 1, we obtain a PAG by running FCI with a significance level of 0.05, then adopt the two strategies to select intervention variables in Stage 2. In Stage 3, we learn the marks with interventional data and orientation rules. We show the results for $d = 10$ in Tab. 1. The MCMC strategy is effective by noting that the number of interventions is smaller compared to the random strategy. According to

Table 1: Number of interventions, normalized SHD, F1 score, number of correctly/ wrongly learned marks by interventions, and number of correctly/wrongly learned marks in PAG over 100 simulations with $d = 10$ and varying p in the format of mean $\pm$ std.

Strategy-p	# int.	Norm. SHD	F1	# correct int.	# wrong int.	# correct PAG	# wrong PAG
Random-0.10	2.88 ± 1.28	0.02 ± 0.04	0.85 ± 0.29	3.92 ± 2.40	0.11 ± 0.53	4.78 ± 3.11	0.39 ± 0.82
MCMC-0.10	2.77 ± 1.19	0.02 ± 0.04	0.85 ± 0.29	4.00 ± 2.40	0.03 ± 0.22		
Random-0.15	3.30 ± 1.15	0.02 ± 0.05	0.91 ± 0.17	5.17 ± 2.62	0.10 ± 0.41	7.21 ± 3.85	0.40 ± 0.92
MCMC-0.15	3.20 ± 1.03	0.02 ± 0.04	0.92 ± 0.16	5.25 ± 2.66	0.02 ± 0.14		
Random-0.20	3.59 ± 1.22	0.04 ± 0.06	0.91 ± 0.15	6.26 ± 2.75	0.19 ± 0.61	9.26 ± 3.94	0.59 ± 1.30
MCMC-0.20	3.42 ± 1.16	0.03 ± 0.06	0.92 ± 0.15	6.38 ± 2.70	0.07 ± 0.33		
Random-0.25	3.47 ± 1.34	0.08 ± 0.14	0.89 ± 0.18	7.08 ± 3.37	0.06 ± 0.34	11.92 ± 4.01	1.59 ± 2.85
MCMC-0.25	3.22 ± 1.19	0.08 ± 0.14	0.89 ± 0.18	7.05 ± 3.39	0.09 ± 0.35		
Random-0.30	3.64 ± 1.32	0.14 ± 0.15	0.83 ± 0.18	7.03 ± 3.33	0.36 ± 1.01	12.59 ± 4.08	2.63 ± 3.19
MCMC-0.30	3.55 ± 1.37	0.14 ± 0.15	0.83 ± 0.18	7.25 ± 3.50	0.14 ± 0.43		

the SHD and F1 score, the active framework can learn the MAG accurately when d and p are not large. The fifth/sixth column shows the number of marks correctly/wrongly learned by interventions, and the seventh/eighth column shows the number of marks correctly/wrongly learned by FCI. They suggest that the marks are learned accurately in Stage 3, while most of the wrong marks are generated in Stage 1. Hence, it implies that the bottleneck of the framework is the first stage.

However, we note although MCMC can reduce the number of interventions, it is inconspicuous. We argue that it is due to the intrinsic hardness posed by the latent variables: the marks are learned mainly based on BK, and the rules cannot reveal much information although they are complete. For example, for an edge $A \circ\!\!-\!\!\circ X$, when the intervention on X does not result in a significant change on A, *only* $A \circ\!\!\rightarrow X$ can be learned and another intervention is necessary to determine the circle at A. We conduct a simulation study to compare the contribution of MCMC strategy to reducing the number of interventions in learning DAGs in causal sufficiency (CS) and learning MAGs in insufficiency settings (CIS). For CS, we generate DAGs with $d - 3$ nodes to ensure the identical number of observed variables in CS and CIS. We learn PAG/CPDAG with the true covariance matrix to avoid the impact of the first stage on the following steps. The ratio of reduced intervention times by MCMC compared to random strategy in CIS/CS is in the second/third column of Tab. 2. It suggests that MCMC has quite different contributions to the reduction of intervention numbers in the two settings. Note the difference is obvious even if d and p are small when the space of MAGs is not large and we can ensure that 1000 samplings of MAGs are sufficient to approximate the uniform distribution of MAGs. It implies that the difference does not result from the samplings in the second stage. Given that the PAG and CPDAG are true and the orientation rules are complete for CIS and CS (Meek rules), the difference is not due to the first and third stages as well. Further, we divide the learned marks by interventions into two parts: those oriented directly by BK, and those learned by the orientation rules. We show the ratio of marks learned by rules to the total learned marks for CIS and CS in Tab. 2. It implies that for CIS, most learned marks are directly from BK, while for CS most learned marks are from applying Meek rules. It further verifies that the inconspicuous reduction in the number of interventions is due to the intrinsic hardness posed by latent variables.

Table 2: The ratio of reduced intervention times by MCMC compared to random strategy and the ratio of marks learned by orientation rules in causal sufficiency (CS) and insufficiency(CIS) setting over 100 simulations with varying d and $p = 0.2$.

d	%diff-CIS	%diff-CS	%rule-CIS	%rule-CS
6	7.07%	21.43%	0.00%	50.98%
8	2.19%	12.50%	1.46%	63.69%
10	5.79%	14.81%	3.59%	67.30%
12	4.21%	25.15%	5.72%	67.89%
14	5.92%	16.50%	6.75%	70.31%
16	2.74%	27.59%	8.49%	74.41%

6 Conclusion

In this paper, we present the sound and complete rules for orienting a PAG with local background knowledge. Based on the theoretical results, we present the first active learning framework for causal discovery in the presence of latent variables, equipped with a maximal entropy criterion based on MH samplings. The experiments suggest that the proposed method can achieve MAG learning with fewer interventions, and the first stage is the most crucial part for the performance of the framework. In the future, we will study the complete rules for orienting a PAG with general BK.

References

[1] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning - problems, methods and evaluation. *SIGKDD Explor.*, 22 (1):18–33, 2020.

[2] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. *arXiv preprint arXiv:2010.06529*, 2020.

[3] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems 30*, pages 656–666, 2017.

[4] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making - the causal explanation formula. In *Proceedings of the Thirty-Second Conference on Artificial Intelligence*, pages 2037–2045, 2018.

[5] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence and 14th Conference on Innovative Applications of Artificial Intelligence*, pages 567–573, 2002.

[6] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference*, pages 1219–1226, 2006.

[7] Sanghack Lee and Elias Bareinboim. Causal effect identifiability under partial-observability. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119. PMLR, 2020.

[8] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[9] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT Press, 2000.

[10] David Maxwell Chickering. Optimal structure identification with greedy search. *c*, 3:507–554, 2002.

[11] Sara Magliacane, Tom Claassen, and Joris M. Mooij. Ancestral causal inference. In *Advances in Neural Information Processing Systems 29, Barcelona, Spain*, pages 4466–4474, 2016.

[12] Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM International Conference on Knowledge Discovery & Data Mining*, pages 1551–1560, 2018.

[13] Daniel Irving Bernstein, Basil Saeed, Chandler Squires, and Caroline Uhler. Ordering-based causal structure learning in the presence of latent variables. In *The 23rd International Conference on Artificial Intelligence and Statistics*, pages 4098–4108, 2020.

[14] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

[15] Ruichu Cai, Jie Qiao, Kun Zhang, Zhenjie Zhang, and Zhifeng Hao. Causal discovery from discrete data using hidden compact representation. In *Advances in Neural Information Processing Systems*, pages 2666–2674, 2018.

[16] Ignavier Ng, Sébastien Lachapelle, Nan Rosemary Ke, and Simon Lacoste-Julien. On the convergence of continuous constrained optimization for structure learning. *CoRR*, abs/2011.11150, 2020.

[17] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[18] Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1347–1353, 2017.

[19] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D. Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21:89:1–89:53, 2020.

[20] Yangbo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9:2523–2547, 2008.

[21] Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.

[22] Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G. Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. In *Advances in Neural Information Processing System*, pages 3195–3203, 2015.

[23] Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16: 2147–2205, 2015.

[24] Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, pages 5824–5833, 2017.

[25] Juan Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. In *Advances in Neural Information Processing Systems*, 2020.

[26] Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In *Advances in Neural Information Processing Systems*, 2020.

[27] Shen Xinpeng, Ma Sisi, Vemuri Prashanthi, Gyorgy Simon, Michael W Weiner, Paul Aisen, Ronald Petersen, Clifford R Jack, Andrew J Saykin, William Jagust, et al. Challenges and opportunities with causal discovery algorithms: Application to alzheimer's pathophysiology. *Scientific Reports (Nature Publisher Group)*, 10(1), 2020.

[28] Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the 11st Annual Conference on Uncertainty in Artificial Intelligence*, pages 403–410, 1995.

[29] Thomas Richardson, Peter Spirtes, et al. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

[30] R Ayesha Ali, Thomas Richardson, Peter Spirtes, and Jiji Zhang. Orientation rules for constructing markov equivalence classes of maximal ancestral graphs. Technical report, Technical Report 476, Dept. of Statistics, University of Washington, 2005.

[31] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

[32] Bryan Andrews. On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge. In *The 23rd International Conference on Artificial Intelligence and Statistics*, pages 4002–4011, 2020.

[33] Hui Zhao, Zhongguo Zheng, and Baijun Liu. On the markov equivalence of maximal ancestral graphs. *Science in China Series A: Mathematics*, 48(4):548–562, 2005.

[34] Marcel Wienöbst, Max Bannach, and Maciej Liskiewicz. Polynomial-time algorithms for counting and sampling markov equivalent dags. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 12198–12206, 2021.

[35] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

[36] Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

[37] Jiji Zhang and Peter Spirtes. A transformational characterization of markov equivalence for directed acyclic graphs with latent variables. In *Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence*, pages 667–674, 2005.

[38] Ayesha R. Ali, Thomas S. Richardson, Peter Spirtes, and Jiji Zhang. Towards characterizing markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence*, pages 10–17, 2005.

[39] Peter Spirtes and Thomas Richardson. A polynomial time algorithm for determining dag equivalence in the presence of latent variables and selection bias. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*, pages 489–500, 1996.

[40] Martin Charles Golumbic. *Algorithmic graph theory and perfect graphs*. Elsevier, 2004.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] It is mainly a theoretical paper.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] It will be publicly available later.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]