Supplementary - Cost-effectively Identifying Causal Effects When Only Response Variable is Observable

Tian-Zuo Wang [1] **Xi-Zhu Wu** [1] **Sheng-Jun Huang** [2] **Zhi-Hua Zhou** [1]

This is the supplementary for **Cost-effectively Identifying Causal Effects When Only Response Variable is Observable**. In Appendix A, we provide a detailed preliminary. Some notations in the proof are not mentioned in the main paper. We illustrate them here. Interventional-faithfulness, which is an important assumption to our method, is given in Appendix B. Appendix C shows the method of estimating the expectation of causal effects by Monte Carlo Methods. All the proofs for the main paper are provided in Appendix D. Appendix E reports some details of our experiments.

Appendix A: Preliminary

In this part, we provide a detailed illustration of the concepts and notations in the main paper and supplementary.

Let $G = (\mathbf{V}, \mathbf{E})$ be a graph consisting the vertices $\mathbf{V}$ and edges $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$. In our context, the vertices represent the features $X_1, X_2, \cdots, X_p$ and the response variable Y. A directed edge is like $V_i \to V_j$ or $V_j \leftarrow V_i$ with an arrow in the edge, while an undirected edge is like $V_j - V_j$. A *directed (undirected) graph* is a graph whose edges are all directed (undirected). A *partially directed graph* may contain both directed and undirected edges. If we remove all the arrowheads from a graph G, we obtain the *skeleton* of the graph G. For two graph $G_1 = (\mathbf{V}_1, \mathbf{E}_1)$ and $G_2 = (\mathbf{V}_2, \mathbf{E}_2)$, G_1 is a *subgraph* of G_2 if $\mathbf{V}_1 \subseteq \mathbf{V}_2$ and $\mathbf{E}_1 \subseteq \mathbf{E}_2$. The *subgraph induced by* $\mathbf{V}'$ in G comprises of the vertices in $\mathbf{V}'$ and all the edges between the vertices in $\mathbf{V}'$. We denote it by $G[\mathbf{V}']$.

For any two vertices V_i and V_j in graph $G = (\mathbf{V}, \mathbf{E})$, if there is an edge between them, they are *adjacent*. The *adjacency set* of a vertice V_i contains all the vertices adjacent to V_i. We denote it by $adj_i(G)$. If $V_i \to V_j$ ($V_j \to V_i$), V_i is a *parent* (*child*) of V_j. A *directed path* from V_i to V_j represents there exists a set of vertices $\{V_{i_1}, V_{i_2}, \cdots, V_{i_k}\} \subset V$ such that $V_i \to V_{i_1}, V_{i_1} \to V_{i_2}, \cdots, V_{i_k} \to V_j$. The *partially directed path* allows for the undirected edge in the set of vertices. If there is a directed path from V_i to V_j (from V_j to V_i), then V_i is an *ancestor* (*descendant*). Every vertice is an ancestor (descendant) of itself. We denote the parents (children, ancestors, descendants) set of vertice V_i in graph G by $\mathrm{Pa}_i[G], \mathrm{Chd}_i[G], \mathrm{Anc}_i[G], \mathrm{Des}_i[G]$. Sometimes, we will simplify them to $\mathrm{Pa}_i, \mathrm{Chd}_i, \mathrm{Anc}_i, \mathrm{Des}_i$. In a partially directed graph G, *siblings* of vertice V_i is the vertice set in which each has an undirected edge with V_i. We denote the *siblings* of vertice V_i by $\mathrm{Sib}_i[G]$ or Sib_i, and denote the undirected edges set of V_i by E_{Sib_i}.

In a graph, there is a *directed (partially) cycle* if there is a *directed (partially) path* starts and ends in the same point. It is worthy to note that there exists at least one directed edge in the partially cycle. If there is no such a directed cycle, the graph is *acyclic*. For any triples of vertices in a graph such that $V_i \to V_k \leftarrow V_j$ and V_i is not adjacent to V_j, they constitute a *v-structure*, in which V_k is a *collider*. The conditional independence can be identified by *d-separation*. V_i and V_j are not d-separated by a vertice set V' in a path $\mathcal{L}$ if and only if for every collider in the path, at least one of its descendant belongs to V' and V' has no non-colliders in the path $\mathcal{L}$. For two variables set V_1 and V_2, if every path from V_1 to V_2 is d-separated by a variable set V', V_1 is *conditional independent* of V_2 given V'. A (partial) causal graph is a *(partially) directed acyclic graph*, which is *DAG*(*PDAG*) for short. The directed edge reflects a causal relation.

If two DAGs share the same conditional independence, they are *Markov equivalent*. Two DAGs are Markov equivalent if and only if they share the same skeleton and same v-structures. The *Markov equivalence class (MEC)* is a set of DAG in

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China [2]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Correspondence to: Sheng-Jun Huang <huangsj@nuaa.edu.cn>, Zhi-Hua Zhou <zhouzh@lamda.nju.edu.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

which each graph is Markov equivalent to others. An *essential graph* is a partially directed acyclic graph, and the edge is $V_i \to V_j$ if and only if in each DAG of MEC the edge is $V_i \to V_j$. A partially directed graph is a *chain graph* if there is no partially cycle (Lauritzen & Richardson, 2002). As shown in (Andersson et al., 1997), the essential graph is a chain graph. After deleting the directed edges in a chain graph, we divide it into a few *chain components* whose variables are connected in an undirected graph.

In this paper, capital and lower-case letters denote random variables and values respectively. In Pearl's *do-calculus* framework (Pearl, 2009), $do(X = x)$ represents intervening on variable X with value x. The causal effect of X_i on Y is denoted by $P(Y|do(X_i))$. We would claim X_i has a causal effect on Y if X_i is an ancestor of Y because an intervention on X will influence Y. Otherwise, we say X_i has *no causal effect* on Y.

A set of variables $\mathbf{Z}$ is called back-door admissible set for (X_i, Y) in a DAG G if no variable in $\mathbf{Z}$ is a descendant of X_i and $\mathbf{Z}$ blocks every path between X_i and Y that contains an arrow into X_i. By definition, Pa_i is one of the back-door admissible sets for (X_i, Y). With back-door admissible set $\mathbf{Z}$ for (X_i, Y), we have

$$P\big(Y|do(X_i = x_i)\big) = \int_{\mathbf{Z}} P(\mathbf{Z})P(Y|X_i = x_i, \mathbf{Z})\,\mathrm{d}\mathbf{Z}. \tag{1}$$

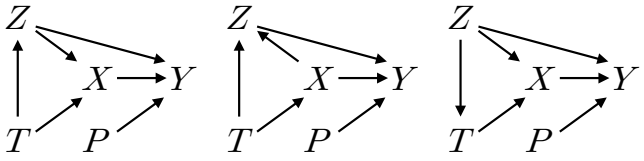
Appendix B: Interventional-faithfulness Assumption

In this section, we give a detailed illustration about the interventional-faithfulness assumption. It is important to our method. We first define *minimal parental back-door admissible set*, followed by the assumption.

Definition 1 (Minimal Parental Back-door Admissible Set (MPS))**.** $\mathbf{M}$ is called a minimal parental back-door admissible set for (X_i, Y) in a DAG G if (1). all variables in $\mathbf{M}$ are parents of X_i, (2). $\mathbf{M}$ is a back-door admissible set for (X_i, Y), (3). no variable in $\mathbf{M}$ is conditional independent of Y given X_i and the other variables in $\mathbf{M}$.

Assumption 1 (Interventional-faithfulness)**.** *For two Markov equivalent DAGs with the same observational distribution, if $X_i \in Anc_Y$ and minimal parental back-door admissible sets for (X_i, Y) are different in the two DAGs, then $P\big(Y|do(X_i = x)\big)$ are different in the two DAGs.*

Now, we give an example to intuitively illustrate it. See Fig 1, the causal effect of $X = x$ on Y is $\int_Z P(Z)P(Y|X, Z)\,\mathrm{d}Z$ in Structure 1, while it is $\int_T P(T)P(Y|X = x, T)\,\mathrm{d}T$ in Structure 2. And they share different minimal parental back-door admissible set $\{Z\}$ and $\{T\}$ for (X, Y). The interventional-faithfulness assumes that given a common joint observational distribution, the causal effect $\int_Z P(Z)P(Y|X, Z)\,\mathrm{d}Z$ is not equivalent to $\int_T P(T)P(Y|X = x, T)\,\mathrm{d}T$, i.e. it is impossible that $P(Y|do(X_i)) = \int_Z P(Z)P(Y|X, Z)\,\mathrm{d}Z = \int_T P(T)P(Y|X = x, T)\,\mathrm{d}T$. The reason we introduce the concept of minimal parental back-door admissible set in our assumption is to differentiate our assumption with "The causal effect implies the only causal graph". For example, the causal effects of X on Y are the same in Structure 1 and 3. But the causal graphs are evidently different. We rule out this situation by introducing minimal parental back-door admissible sets. The assumptions are towards only the causal graphs with different minimal parental back-door admissible sets.



(a) Structure 1 (b) Structure 2 (c) Structure 3

Figure 1. We assume the causal effects of X on Y are different in the first two structures under a common joint distribution, while those in Structure 1 and 3 are equivalent.

Then, we give a brief analysis to show intervention-faithfulness is hardly violated under the faithfulness assumption.

Analysis. *If intervention-faithfulness is violated, that is, for two DAGs G and H with different MPS $\mathbf{V}_1$ and $\mathbf{V}_2$ for (X_i, Y),*

their causal effects are the same. It holds

$$P_G(Y|do(X=x)) = \int P(\mathbf{V}_1)P(Y|X=x,\mathbf{V}_1)\,\mathrm{d}\mathbf{V}_1, \tag{2}$$

$$P_H(Y|do(X=x)) = \int P(\mathbf{V}_2)P(Y|X=x,\mathbf{V}_2)\,\mathrm{d}\mathbf{V}_2. \tag{3}$$

We find $\mathcal{A}$ and $\mathcal{B}$ such that $\mathbf{V}_1 \cup \mathcal{A} = \mathbf{V}_2 \cup \mathcal{B}$, $\mathcal{A} \subseteq \mathbf{V}_2, \mathcal{B} \subseteq \mathbf{V}_1, \mathcal{A} \cap \mathcal{B} = \emptyset$. If the causal effects are the same, that is

$$\begin{aligned}&\int P(\mathbf{V}_1)P(Y|X=x,\mathbf{V}_1,\mathcal{A})P(\mathcal{A}|X=x,\mathbf{V}_1)\,\mathrm{d}\mathbf{V}_1\,\mathrm{d}\mathcal{A}\\ =&\int P(\mathbf{V}_2)P(Y|X=x,\mathbf{V}_2,\mathcal{B})P(\mathcal{B}|X=x,\mathbf{V}_2)\,\mathrm{d}\mathbf{V}_2\,\mathrm{d}\mathcal{B}.\end{aligned}$$

We thus have

$$\begin{aligned}0 =& \int \big[P(\mathbf{V}_1)P(\mathcal{A}|X=x,\mathbf{V}_1) - P(\mathbf{V}_2)P(Y|X=x,\mathbf{V}_2,\mathcal{B})P(\mathcal{B}|X=x,\mathbf{V}_2)\big]P(Y|X=x,\mathbf{V}_1,\mathcal{A})\,\mathrm{d}\mathbf{V}_1\,\mathrm{d}\mathcal{A}\\ =& \int f(\mathcal{A},\mathbf{V}_1)P(Y|X=x,\mathbf{V}_1,\mathcal{A})\,\mathrm{d}\mathbf{V}_1\,\mathrm{d}\mathcal{A}, \forall x,\end{aligned}$$

where $f(\mathcal{A},\mathbf{V}_1) = P(\mathbf{V}_1)P(\mathcal{A}|X=x,\mathbf{V}_1) - P(\mathbf{V}_2)P(Y|X=x,\mathbf{V}_2,\mathcal{B})P(\mathcal{B}|X=x,\mathbf{V}_2)$. Here it is not hard to see

$$\int f(\mathcal{A},\mathbf{V}_1)\,\mathrm{d}\mathbf{V}_1\,\mathrm{d}\mathcal{A} = 0.$$

We first prove that $f(\mathcal{A},\mathbf{V}_1) \not\equiv 0$. If $f(\mathcal{A},\mathbf{V}_1) \equiv 0$, that means

$$\begin{aligned}P(\mathbf{V}_1)P(\mathcal{A}|X=x,\mathbf{V}_1) &= P(\mathbf{V}_2)P(\mathcal{B}|X=x,\mathbf{V}_2)\\ P(X=x|\mathbf{V}_1) &= P(X=x|\mathbf{V}_2).\end{aligned}$$

Without loss of generality, we assume $\mathcal{A} \neq \emptyset$. We conclude

$$P(X=x|\mathbf{V}_1) = P(X=x|\{\mathbf{V}_2\backslash\mathcal{A}\},\mathcal{A}) = P(X=x|\{\mathbf{V}_2\backslash\mathcal{A}\}).$$

The second equation holds because $\mathcal{A} \cap \mathbf{V}_1 = \emptyset$. Hence $X \perp \mathcal{A}|\{\mathbf{V}_2\backslash\mathcal{A}\}$. It contradicts the faithfulness assumption, because there exists edges between X and variables in $\mathcal{A}$.

Hence, we know that $f(\mathcal{A},\mathbf{V}_1) \not\equiv 0$. We notice if the interventional-faithfulness is violated, that means given any x, the weighted sum of a series of distribution of Y should happen to equal to zero, where the sum of the weight is zero, and the weight is not a function of Y as well as not always zero. It is reasonable to think such a situation hardly happens.

Appendix C: The Expectation Estimation of Causal Effects

In this part, we introduce our method for estimating the expectation of causal effects in the main paper. The equation is

$$\mathbb{E}\big(Y|do(X_i=x_i)\big) = \int_Y Y \int_{\mathbf{M}_j} P(\mathbf{M}_j)P(Y|X_i=x_i,\mathbf{M}_j)\,\mathrm{d}\mathbf{M}_j\,\mathrm{d}Y, \tag{4}$$

$$= \int_{\mathbf{M}_j} P(\mathbf{M}_j)\mathbb{E}(Y|X_i=x_i,\mathbf{M}_j)\,\mathrm{d}\mathbf{M}_j. \tag{5}$$

The estimation in (4) is prone to suffer the curse of dimensionality, because it gets harder to estimate $P(Y|X_i=x_i,\mathbf{M}_j)$ as the growth of the dimension of $\mathbf{M}_j$. Hence we estimate it avoiding the high-dimensional estimation by Monte Carlo Methods. We divide the estimation of (5) into two steps. First, we train a regression model $\hat{f}$ from X_i and $\mathbf{M}_j$ to Y based on the observational data in order to predict the expectation $\mathbb{E}(Y|X_i=x_i,\mathbf{M}_j)$. The predicted value for $X_i=x_i$ and $\mathbf{M}_j=\mathbf{m}_0$ is $\hat{f}(x_i,\mathbf{m}_0)$. Then, we sample $\mathbf{m} \sim P(\mathbf{M}_j)$ with replacement from the observational data. With a sample $\mathbf{m}_k$

and the intervention value x_i, we estimate $\mathbb{E}(Y|X = x_i, \mathbf{m}_k)$ with the trained regression model. The result is denoted by $\hat{Y}_k$. Repeat the process of sampling and estimation for many times and get $\hat{Y}_1, \hat{Y}_2, \cdots, \hat{Y}_n$. Then,

$$\hat{\mathbb{E}}_{G_j}\big(Y|do(X = x_i)\big) = \frac{1}{n}\sum_{k=1}^{n}\hat{Y}_k. \tag{6}$$

Here we give an example to provide reasonableness for the regression step in estimating the expectation $\mathbb{E}(Y|X_i = x_i, \mathbf{M}_j)$. Assume a linear relation $Y = \beta_0 X + \beta^{\mathrm{T}}\mathbf{M}_j + \epsilon, \epsilon \sim (0, \sigma^2)$ between the variables, where $\mathbf{M}_j$ is the minimal parental back-door admissible set. It holds that $\mathbb{E}(Y|X = x_i, \mathbf{m}_0) = \beta_0 x_i + \beta^{\mathrm{T}}\mathbf{m}_0$. Our predicted value is $\hat{f}(x_i, \mathbf{m}_0) = \hat{\beta}_0 x_i + \hat{\beta}^{\mathrm{T}}\mathbf{m}_0$. Due to the property of linear model, $\hat{\beta}_0$ and $\hat{\beta}$ are all unbiased estimate of β_0 and β. Hence, $\hat{f}(x_i, \mathbf{m}_0)$ is an unbiased estimate of $\mathbb{E}(Y|X = x_i, \mathbf{m}_0)$. It seems reasonable to use the regression result to estimate the expectation given x_i and $\mathbf{m}_0$.

Appendix D: Theoretical Guarantee

In this part we provide the proofs in the main paper. In Part D.1, we present some detailed proofs about the proposed method, which is in Section 3 in the main paper. In Part D.2, the related proofs about the causal effect identifiability are given. In Part D.3, we give the proof about the intervention cost analysis.

D.1: proof about the proposed method

Theorem 1. *The causal effect of each variable on the response variable Y is identifiable if and only if all ancestor edges (ancestor causal structure) are identified.*

Proof. The adequacy and necessity are proved, respectively.
$\Rightarrow$. We prove it by reduction to absurdity. Without loss of generality, we assume an ancestor edge $X_i \to X_j$ has not been identified. According to the definition of *ancestor edges*, after deleting the edge between X_i and X_j, there exists at least one variable from X_i, X_j having a directed path to Y. If both have directed paths to Y (the edge between X_i and X_j is not in the path), whether $X_i \to X_j$ or $X_i \leftarrow X_j$ determines different minimal parental back-door variable sets for (X_i, Y) in (1), which takes a different causal effect of X_i on Y in general. If only one variable has a directed path to Y while the other one is not located in the path, we assume it is $X_i \to \cdots \to Y$. If $X_i \to X_j$, X_j has no causal effect to Y. If $X_i \leftarrow X_j$, intervention on X_j will take a change to Y. Hence, we cannot determine the causal effect of X_j on Y without the direction of the $X_i - X_j$, which contradicts the condition.

$\Leftarrow$. If the ancestors edges are all identified, the set Anc_Y is determined. For any $X \in V \backslash \mathrm{Anc}_Y$, it holds $\mathbb{E}(Y|do(X = x)) = \mathbb{E}(Y)$. For $X \in \mathrm{Anc}_Y$, because all edges of X are identified, we can determine the back-door variables and calculate the causal effect by back-door criterion. Hence, the causal effect of each variable on Y is identifiable. □

Proposition 2. *In a chain component C of chain graph G, if the response variable $Y \notin C$ and there does not exist a directed path from v to Y for any variable $v \in C$ in G, then there is no directed path from C to Y in the causal graph.*

Proof. If such a path $v \to \cdots \to Y$ exists in the causal graph, then $v \to V_1 \to \cdots \to V_k \to Y$ and some edges $V_i \to V_{i+t} \in G_\tau$, $V_{i-1} \to V_i \notin G_\tau$ and $V_{i+t} \to V_{i+t+1} \notin G_\tau$, where G_τ is a chain component. According to Lemma 10 by He & Geng (2008), if a variable S outside a chain component $\mathcal{G}'$ has a directed edge to X in $\mathcal{G}'$, then there is a directed edge from S to all the variables in $\mathcal{G}'$. Therefore, the path can be shorten to $v \to V_1 \to V_{i-1} \to V_{i+t} \to \cdots \to V_k \to Y$. Repeat it and we can get a directed path from a variable in C to Y that doesn't contain any edges in other chain components. It contradicts the condition. □

D.2 proof about causal effect identifiability

Lemma 1. *Under Assumption 1 and our intervention variable selection strategy, if X is intervened in chain component C, $Y \notin C$, undirected edges exist between $\mathrm{Anc}_Y[\ddot{C}]$, then all the undirected edges between X and the "next" variable located in the shortest undirected path $\mathcal{P}$ from X to any variable $Z \in \mathrm{Anc}_Y[\ddot{C}]$ can be identified by the intervention, the "next" variable is the one adjacent to X in $\mathcal{P}$.*

Proof. For any other variable $Z(\neq X)$ in $\text{Anc}_Y[\ddot{C}]$, we assume the shortest undirected path between X and Z is $X - V_{a_1} - V_{a_2} - \cdots - V_{a_k} - Z$. We prove the causal relation between X and V_{a_1} can be identified by intervening on X because the different orientations of $X - V_{a_1}$ lead to different minimal parental back-door admissible set for X, Y, which lead to different causal effects by interventional-faithfulness assumption.

If $X \to V_{a_1}$, V_{a_1} cannot be in the minimal parental back-door admissible set $\mathbf{M}_X$ for (X, Y).

While if $V_{a_1} \to X$, V_{a_1} must be in $\mathbf{M}_X$. Otherwise, there exists a $\mathbf{T} \subseteq \text{Sib}_X \bigcup \text{Pa}_X$, such that $V_{a_1} \perp Y | \{X, \mathbf{T}\}$. With $Z \in \text{Anc}_Y[\ddot{C}]$, we first prove $\{X, \text{Pa}_X\}$ cannot d-separate V_{a_1} and Y in the path $X - V_{a_1} - V_{a_2} - \cdots - V_{a_k} - Z \to \cdots \to Y$:

At first, we notice $\{X, \text{Pa}_X\}$ cannot d-separate the shortest path from V_{a_1} to Z. To achieve d-separation $V_{a_1} \perp Y | \{X, \text{Pa}_X\}$, at least one variable V' should d-separate the directed path from Z to Y. In the condition that $V' \in \text{Pa}_X$, if $V' \notin C$, it holds $V' \to Z$ (Lemma 10 of He & Geng (2008)), hence the d-separation is impossible. If $V' \in C$, if the d-separation holds, all the directed path from Z to Y will go through such V', which contradicts the definition of $\text{Anc}_Y[\ddot{C}]$, because $Z \in \text{Anc}_Y[\ddot{C}]$ so that Z should have a directed path to Y in which no variables in C exist. Hence the variables in PA_X cannot d-separate the path from Z to Y;

We also notice X is evidently not in the directed path from Z to Y, which leads to it cannot d-separate the path from Z to Y. We then get the conclusion $\{X, \text{Pa}_X\}$ cannot d-separate the path from V_{a_1} to Y.

Therefore, if $V_{a_1} \perp Y | \{X, \mathbf{T}\}$, there is at least one $V_t \in \text{Sib}_X$ such that V_{a_1} is d-separated with Y by V_t in that path. We have $V_t \in \{V_{a_2}, \cdots, V_{a_k}\}$. There is an undirected path $X - V_t - \cdots - Z$, which contradicts the shortest path assumption. Therefore V_{a_1} must be in $\mathbf{M}_X$. By interventional-faithfulness assumption, the causal effects when $X \to V_{a_1}$ and $X \leftarrow V_{a_1}$ are different, so that we can identify this edge with the distribution of Y under intervention on X. □

D.3 proof about intervention cost analysis

In this part, we provide the related proofs for the interventional cost analysis section in the main paper. We assume plenty of observational data. According to them we can obtain the correct essential graph. Based on the essential graph, we make experiments and analyze the expected number of interventions to identify the ancestor causal structure, i.e., make causal effect identification. The number of interventions do not include of that obtaining the observational data.

Proposition 4. *For a line skeleton with $p + 1 \geq 4$ variables $X_1, \cdots, X_p, Y$, if all the causal relations and positions of variables $X_1, \cdots, X_p, Y$ are totally random, then the expected number of interventions to make causal effect identification is $\frac{19}{8} - \frac{39}{8p+8} + \frac{6}{p+1}(\frac{1}{2})^p < 3$.*

Proof. At first, we consider Y is at one end of a line with m variables and denote the intervention times in such line by $\#s(m)$. For the graph $Y - X_1 - X_2 - \cdots$, there are three conditions $Y \leftarrow X_1 - \cdots$, $Y \to X_1 \leftarrow X_2 - \cdots$, $Y \to X_1 \to X_2 - \cdots$ with possibilities $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{4}$. For the second condition, they form a v-structure so that we can identify no variables from $X_1 \cdots, X_{m-1}$ are the ancestors of Y by observational data. Hence the number of interventions is zero. For others condition, we will intervene on X_1 to identify the edges between X_1 and Y. We keep intervening until identifying all ancestor edges. We have

$$\mathbb{E}(\#s(m)) = 0 \times \frac{1}{4} + 1 \times \frac{1}{4} + (1 + \mathbb{E}(\#s(m-1))) \times \frac{1}{2}.$$

Recursively, we can conclude that

$$\mathbb{E}(\#s(m)) = \frac{2}{3} - (\frac{1}{2})^{m-1}, m \geq 2.$$

When we analyze the number of interventions in a random line skeleton, the randomness is generated from three parts. The first one is the random positions of Y. The second one is the random causal edges in the real structure. The last one is the intervention variable selection. Y is located in each position of the line skeleton with the same possibility. Without loss of generality, we assume the skeleton is $X_1 - X_{i-1} - Y - X_i - \cdots - X_p, 3 \leq i \leq p-1$ (The condition that $i = 1, 2, p, p+1$ can be analyzed similarly, we thus omit them here). We first consider the edges $X_{i-1} - Y - X_i$. In the real causal graph, there are four conditions $X_{i-1} \leftarrow Y \to X_i$, $X_{i-1} \to Y \leftarrow X_i$, $X_{i-1} \to Y \to X_i$, $X_{i-1} \leftarrow Y \leftarrow X_i$, each of them is with the possibility $\frac{1}{4}$.

Take $X_{i-1} \to Y \to X_i$ as an example, if the edges between X_i and X_{i+1} is $X_i \leftarrow X_{i+1}$, the edge between Y and X_i can be identified by observational data because Y, X_i, X_{i+1} form a v-structure, so that we do not need to do interventions to identify $X_i - X_{i+1}$. And the number of interventions to identify $X_{i-1} - Y$ is 1. When $X_i \to X_{i+1}$, we select the intervention variable from X_{i-1} and X_i with the same possibility $\frac{1}{2}$. If we intervene on X_{i-1}, both $X_{i-1} \to Y$ and $Y \to X_i$ can be identified because of Meek rules, that is we take one experiment to identify $X_i \to Y \to X_{i+1}$. But if we intervene on X_i, we have to take another intervention on X_{i-1} to identify the edge $X_{i-1} - Y$, that is we take two experiments to identify $X_i \to Y \to X_{i+1}$. Because the variables after X_{i+1} can not be ancestors of Y, we do not identify their causal relation. The expected number of interventions to identify the causal effect of the variables in $X_1 - \cdots - X_{i-1}$ on Y is $\mathbb{E}(\#s(i-1))$.

Hence, when Y is in the i-th position of the line and $X_{i-1} \to Y \to X_i$, the expected number of experiments to identify the causal effect of each variable on Y is

$$\begin{aligned}\mathbb{E}_{X_{i-1}\to Y\to X_i}(\#(i)) =& \frac{1}{2}\times 1 + \frac{1}{2}\times(\frac{1}{2}\times 1 + \frac{1}{2}\times 2) + \mathbb{E}(\#s(i-1)) \\ =& \frac{5}{4} + \mathbb{E}(\#s(i-1)).\end{aligned}$$

Similarly, we have the expected number of experiments when $X_{i-1} \leftarrow Y \to X_i$, $X_{i-1} \to Y \leftarrow X_i$, $X_{i-1} \leftarrow Y \leftarrow X_i$. In total, we have

$$\begin{aligned}\mathbb{E}(\#\mathcal{E}(p+1)) =& \frac{1}{p+1}\sum_{i=1}^{p+1}\frac{1}{4}\big(\mathbb{E}_{X_{i-1}\to Y\to X_i}(\#(i)) + \mathbb{E}_{X_{i-1}\to Y\leftarrow X_i}(\#(i)) \\ &+ \mathbb{E}_{X_{i-1}\leftarrow Y\to X_i}(\#(i)) + \mathbb{E}_{X_{i-1}\leftarrow Y\leftarrow X_i}(\#(i))\big) \\ =& \frac{19}{8} - \frac{39}{8p+8} + \frac{6}{p+1}(\frac{1}{2})^p, p \geq 3,\end{aligned}$$

i is the position of Y in the line skeleton. The proof completes. □

By Eberhardt (2007), the expected number of interventions have been studied in depth, where they take singleton hard interventions and observe whole variables, as well as discover the edges by whether the distribution of some variable takes a change under an intervention on other variables. In the following, instead of repeating the setting and approach above, we will say "by approach of Eberhardt (2007)" for short. We compare the intervention cost by our approach with them. But our approach is only allowed to have the interventional data of response variable. Because we focus on making causal effect identification, i.e. discovering ancestor causal structure, while Eberhardt (2007) identifies the whole causal structure, we first give following lemmas to conclude the properties of their approach when focusing on ancestor causal structure.

Definition 2 (Causal order)**.** Given a set of variables $\{X_1, \cdots, X_p, Y\}$ and their causal graph, its causal order is an order that any variable is not the ancestor of any former variables in the causal order.

Lemma 2 (Eberhardt (2007))**.** *Given a set of $p+1 \geq 2$ causally sufficient variables, the worst case expected number of experiments necessary and sufficient to discover the causal structure is $\frac{2}{3}(p+1) - \frac{1}{3}$ experiments if only one variable can be subject to a structural intervention per experiment.*

The worst case is when the causal graph is complete. That is, for a complete causal graph, the expected number of interventions is $\frac{2}{3}(p+1) - \frac{1}{3}$. Then, we conclude Lemma 3.

Lemma 3. *For a complete causal graph G with variables $X_1, X_2 \cdots, X_p, Y$, in which Y is the descendant of all the other variables, the expected number of interventions to identify the ancestor causal structure is $\frac{2}{3}(p+1) - \frac{1}{3}$ by the approach of Eberhardt (2007), i.e. taking singleton hard interventions and observing the whole variables ,. The ratio of the expected number to the variable number converges to $\frac{2}{3}$ when $p \to \infty$.*

In this condition, all the edges in graph G are ancestor edges. We thus have to identify all of them. Hence the expected number is same as Lemma 2. Next, we will consider the expected number of interventions to identify the ancestor causal structure, i.e. make causal effect identification, in general complete causal graph.

Lemma 4. *The expected number of interventions $T(p)$ to identify the ancestor causal structure of a random complete causal graph composed of $X_1, \cdots, X_p, Y$ by the approach of Eberhardt (2007), i.e. taking singleton hard interventions and observing the whole variables, is bounded by $\frac{p+1}{3} + \frac{2}{3}\frac{1}{p+1} \leq T(p) \leq \frac{p+1}{3} + \frac{5}{3}\frac{1}{p+1} + \frac{p+2}{p+1}\ln\frac{p+1}{2}$. The ratio of the expected number to the variable number converges to $\frac{1}{3}$ when $p \to \infty$.*

Proof. In a complete graph, there is an exact causal order for $p+1$ variables, in which Y is located in each position with the same possibility. Without loss of generality, we assume the causal order is $X_1, X_2, \cdots, X_{i-1}, Y, X_i, \cdots, X_p$. We divide the graph into two parts. One is the subgraph induced by $Y, X_j, 1 \leq j \leq i-1$. The expected number of interventions is $S(i) = \frac{2}{3}i - \frac{1}{3}, i \geq 2$ by Lemma 3 because Y is the descendant of all other variables $X_1, \cdots, X_{i-1}$. The other is the subgraph induced by $Y, X_j, j \geq i$.

First, we prove for a complete causal graph with m variables in which Y is not descendants of any variables, the expected number of interventions is $1 + \sum_{i=3}^{m} \frac{1}{i}, m \geq 3$. We denote it by $L(m)$. We assume the variables are $X_1, \cdots, X_{m-1}, Y$ (It is different from the notation in the beginning). And we assume the causal order is $X_0, X_1, \cdots, X_{m-1}$, where X_0 is Y. Because each variable is equally likely to be subject to an intervention in the first experiment. We denote the intervention variable by X_j. It holds

$$L(m) = 1 + \frac{1}{m}\sum_{j=0}^{m-1} L(j), m \geq 2,$$

where $L(0) = 0, L(1) = 0, L(2) = 1$. The first term 1 is an intervention cost on X_j. By the interventional data of the full variables, we can see $X_k \to X_j, 0 \leq k < j$ and $X_k \leftarrow X_j, k > j$. Hence we identify that $X_k, k \geq j$ are the descendants of X_j, X_0 is an ancestor of X_j. We thus know all of $X_k, k \geq j$ are descendants of X_0 and their undirected edges are not ancestor edges. The remaining undirected edges to be oriented are just that between $X_0, X_1, \cdots, X_{j-1}$. The expected number of interventions to discover these is $L(j)$. By transformations,

$$\sum_{j=0}^{m-1} [L(m) - L(j)] = m.$$

We denote $L(m) - L(j) = \sum_{i=j+1}^{m} b_i$. It holds

$$\sum_{i=1}^{m} i b_i = m,$$

Similarly, we have

$$\sum_{i=1}^{m+1} i b_i = m + 1,$$

Calculate the difference between the two equations above. It concludes

$$b_i = \frac{1}{i}, i \geq 3,$$
$$L(m) = 1 + \sum_{i=3}^{m} \frac{1}{i}.$$

For the complete random causal graph with variables $X_1, X_2, \cdots, X_p, Y$. We denote the expected number of interventions to discover the ancestor causal structure by $T(p)$. It should be less than the sum of expected number $S(i)$ to identify the subgraph induced by $X_1, X_2, \cdots, X_{i-1}, Y$ and the expected number $L(p+2-i)$ to identify the subgraph induced by

$Y, X_i, X_{i+1}, \cdots, X_p$. Hence we have

$$\begin{aligned}
T(p) &\leq \frac{1}{p+1}\sum_{i=1}^{p+1}(S(i)+L(p+2-i)) \\
&= \frac{1}{p+1}\sum_{i=1}^{p+1}S(i) + \frac{1}{p+1}\sum_{i=1}^{p+1}L(i) \\
&= (\frac{p+1}{3}+\frac{2}{3}\frac{1}{p+1}) + (\frac{1}{p+1}+\frac{p+2}{p+1}\sum_{j=3}^{p+1}\frac{1}{j}) \\
&\leq \frac{p+1}{3}+\frac{5}{3}\frac{1}{p+1}+\frac{p+2}{p+1}\int_{i=2}^{p+1}\frac{1}{j}\,\mathrm{d}j \\
&= \frac{p+1}{3}+\frac{5}{3}\frac{1}{p+1}+\frac{p+2}{p+1}\ln\frac{p+1}{2}.
\end{aligned}$$

And then, it is evident that

$$\begin{aligned}
T(p) &\geq \frac{1}{p+1}\sum_{i=1}^{p+1}S(i) \\
&= \frac{p+1}{3}+\frac{2}{3}\frac{1}{p+1}.
\end{aligned}$$

We then get the desired conclusion. □

Next, we provide the expected number of interventions by our approach. We emphasize only the interventional data of response variable is observable in our approach.

Lemma 5. *For a complete causal graph G with variables $X_1, X_2 \cdots, X_p, Y, p \geq 3$, in which Y is the descendant of all other variables. Given the essential graph Ess_G (the essential graph is the skeleton when the causal graph is complete), the expected number of interventions $F(p)$ to discover the ancestor causal structure by our approach is bounded by $\frac{2}{3}(p+1)-1 \leq F(p) \leq \frac{2}{3}(p+1)-\frac{2}{3}+\ln\frac{p}{2}$, in the setting that only the interventional data of response variable is observable.*

Proof. The complete graph has an exact causal order. Without loss of generality, we assume the order is $X_1, X_2, \cdots, X_p, Y$. According to the condition, each variable from $X_1, \cdots, X_p$ is equally likely to be subject to an intervention in the first experiment.

We consider an intervention on X_i. We can identify the edge $X_i \to Y$. Besides, if X_j is such that $X_j \to X_i, X_j \to Y$, the two edges can be identified. It is because these back-door paths cannot be d-separated by any other variables, the variable X_j must exist in all minimal parental back-door admissible sets of X_i consistent to the real causal structure. Hence we know some edges of $X_j, j \leq i$ can be identified by intervention on X_i, while no edges of $X_j, j > i$ can be identified. Evidently all variables belong to a common chain component. In our intervention variable selection criterion, we will select the variable with the maximum siblings. Hence in the next intervention we will select a variable from $X_{i+1}, \cdots, X_p$ with the same possibility. We repeat the process until intervening on X_p, because no edges of X_p can be identified unless we intervene on it and it has the maximum siblings. This is the first stage ending with intervening on X_p. The expected number of interventions in the first stage is denoted by $K(p)$. $K(1) = 1$. We have

$$\begin{aligned}
K(p) &= 1 + \frac{1}{p}\sum_{i=1}^{p}K(p-i) \\
&= 1 + \frac{1}{p}\sum_{i=0}^{p-1}K(i).
\end{aligned}$$

Similar to the derivation process in Lemma 4, we conclude that

$$K(p) = 1 + \sum_{i=3}^{p} \frac{1}{i}.$$

After we intervene on X_p, we have identified all the edges $X_i \to Y, 1 \leq i \leq p$ and $X_i \to X_p, i < p$. Hence the remaining undirected edges are a subset of edges between $X_1, \cdots, X_p$. Then we consider the expected number of interventions to identify the causal structure between $X_1, X_2, \cdots, X_p$. In the first stage, it is possible that we have intervened on some variables from $X_1, \cdots, X_{p-1}$. Here we just consider the upper bound of intervention cost, that is we consider no edges between $X_1, \cdots, X_{p-1}$ have been identified. In this condition, when we intervene on some variable X_k, all the other variables from $X_1, \cdots, X_{p-1}$ are the "next" variable in the shortest path to Y. By Lemma 1, the edges between X_k and X_m can be identified for all $1 \leq m \leq p-1, m \neq k$. Hence in this condition identifying the causal edges from the interventional data of Y by our approach is equivalent to that from the interventional data of full variables in Eberhardt (2007). The expected number $S(i-1)$ of interventions to discover the subgraph with $i-1$ variables is $\frac{2}{3}(i-1) - \frac{1}{3}$, as shown in Lemma 2.

We denote the expected number of interventions to identify the ancestor causal structure in the complete graph with $p+1$ variables in which Y is the descendant of all other variables by $F(p)$. According to the analysis above, the total intervention process is divided into two stages. The first one is we intervene until intervening on X_p. The expected number of interventions is $K(p)$. The other is we identify the causal structure between $X_1, \cdots, X_{p-1}$. Hence we have

$$\begin{aligned} F(p) &\leq K(p) + S(p-1), p \geq 3 \\ &\leq 1 + \sum_{i=3}^{p} \frac{1}{i} + \frac{2}{3}(p-1) - \frac{1}{3}, p \geq 3 \\ &\leq \frac{2}{3}p + \ln \frac{p}{2}, p \geq 3. \end{aligned}$$

We thus obtain an upper bound for the expected number of interventions to identify the ancestor causal structure. Then, we present a lower bound. For the skeleton Ess of the complete causal graph, we orient all the edges $X_i \to Y, 1 \leq i \leq p$ and get a new PDAG H. It is evident that the expected number of interventions to identify the ancestor causal structure based on H is less than that to identify the ancestor causal structure based on the Ess. Because the intervention strategy in these two graphs are the same, but the former one has less edges to be oriented. Hence a lower bound for $F(p)$ is the expected number of interventions to identify the ancestor causal structure based on H. We have shown that in this condition identifying the causal edges from the interventional data of Y by our approach is equivalent to that from the interventional data of full variables in Eberhardt (2007). Hence the number if $\frac{2}{3}p - \frac{1}{3}$. It thus holds

$$\begin{aligned} \frac{2}{3}p - \frac{1}{3} \leq F(p) \leq \frac{2}{3}p + \ln \frac{p}{2}, p \geq 3, \\ \frac{2}{3} - \frac{1}{p+1} \leq \frac{F(p)}{p+1} \leq \frac{2}{3} - \frac{2}{3p+3} + \frac{1}{p+1} \ln \frac{p}{2}, p \geq 3, \\ \lim_{p \to \infty} \frac{F(p)}{p+1} = \frac{2}{3}. \end{aligned}$$

The proof completes. □

Proposition 5. *For a complete skeleton with $p+1 \geq 4$ variables $X_1, \cdots, X_p, Y$, if all the causal relations and positions of variables $X_1, \cdots, X_p, Y$ are totally random, then the expected number of interventions to make causal effect identification is less than $\frac{5}{6}(p+1) - \frac{11p-10}{6p+6} + \ln \frac{p}{2}$.*

Proof. In a complete graph, there is an exact causal order for $p+1$ variables, in which Y is located in each position with the same possibility. Without loss of generality, we assume the causal order is $X_1, X_2, \cdots, X_i, Y, X_{i+1}, \cdots, X_p$. Each variable X_j after Y costs one intervention to identify its edge with Y. (Because when we intervene on $X_j, j \leq i$, these edges cannot be identified. And when we intervene on $X_j, j > i$, the distribution of Y remains. The edge between other $X_k, k > i, k \neq j$ and Y is still unidentified.) There thus needs $p - i$ interventions to discover these edges $(X_j - Y, j > i)$. For the complete

subgraph induced by Y and all its ancestors $X_1, \cdots, X_i, Y$, we denote the expected number of interventions by $F(i)$. Hence the expected number $C(p)$ to identify the whole ancestor causal structure meets

$$\begin{aligned} C(p) &= \frac{1}{p+1}\sum_{i=0}^{p}(p-i+F(i)) \\ &= \frac{p}{2} + \frac{1}{p+1}(1+\frac{3}{2}+\sum_{i=3}^{p}(F(i))). \end{aligned}$$

Lemma 5 provides the upper bound for the expected number of interventions $F(i)$ to identify the causal structure between $X_1, X_2, \cdots, X_i, Y$. Hence

$$\begin{aligned} \sum_{i=3}^{p} F(i) &\leq \sum_{i=3}^{p}(\frac{2}{3}i + \sum_{j=3}^{i}\frac{1}{j}) \\ &\leq \frac{1}{3}p^2 + \frac{p}{3} - 2 + \sum_{j=3}^{p}\sum_{i=j}^{p}\frac{1}{j} \\ &\leq \frac{1}{3}p^2 + \frac{p}{3} - 2 + \sum_{j=3}^{p}(p-j+1)\frac{1}{j} \\ &\leq \frac{1}{3}p^2 + \frac{p}{3} - 2 - (p-2) + (p+1)\sum_{j=3}^{p}\frac{1}{j} \\ &\leq \frac{1}{3}p^2 - \frac{2p}{3} + (p+1)\int_{j=2}^{p}\frac{1}{j}\,\mathrm{d}j \\ &\leq \frac{1}{3}p^2 - \frac{2p}{3} + (p+1)\ln\frac{p}{2}. \end{aligned}$$

Hence,

$$\begin{aligned} C(p) &\leq \frac{p}{2} + \frac{5}{2(p+1)} + \frac{1}{p+1}(\frac{1}{3}p^2 - \frac{2p}{3} + (p+1)\ln\frac{p}{2}) \\ &\leq \frac{5}{6}(p+1) - \frac{11p-10}{6p+6} + \ln\frac{p}{2}, p \geq 3. \end{aligned}$$

The proof completes. □

Appendix E: Supplement to the experiments

In this part, we give a detailed illustration about the experiments in the main paper. In the example to show the process of identifying causal structure by various approaches, we generate the non-linear data by the following equations. And when

we intervene, the intervention value is set to the mean of the intervened variable in the observational data.

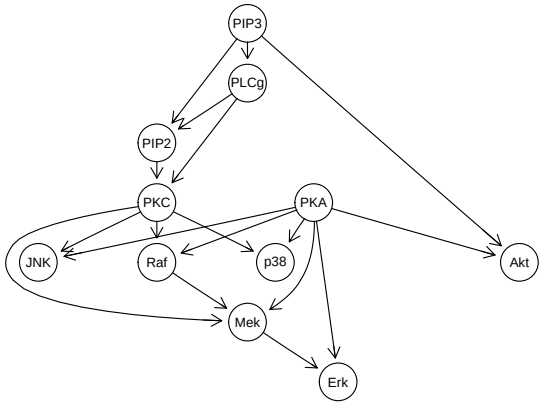
$$
\begin{aligned}
X_1 &= E_1, & E_1 &\sim N(0,1),\\
X_2 &= \frac{3}{2}\sin(X_1) + E_2, & E_2 &\sim U(-0.2, 0.2),\\
X_3 &= -2X_1 + X_2^2 + E_3, & E_3 &\sim N(0, 0.5),\\
X_4 &= 2X_3 + E_4, & E_4 &\sim U(-0.01, 0.01),\\
X_5 &= (X_4 + E_5)^2, & E_5 &\sim N(0, 0.3),\\
X_6 &= \exp^{\sin(X_5+2X_2)+2} + E_6, & E_6 &\sim t(10),\\
X_7 &= \tan(X_4) + E_7, & E_7 &\sim \text{Exp}(2),\\
X_8 &= (\frac{1}{1+X_7})^{\frac{2}{3}} + E_8, & E_8 &\sim \chi^2(4),\\
X_9 &= X_7^2 + X_8 + E_9, & E_9 &\sim N(0, 0.5),\\
X_{10} &= 2X_6 + E_{10}, & E_{10} &\sim \chi^2(2),\\
X_{11} &= \frac{1}{X_6} + X_{10} + E_{11}, & E_{11} &\sim \text{Exp}(4).
\end{aligned}
$$

In the application to the real-world data, we apply our approach on a dataset used in causal discovery with both observational and interventional data (Sachs et al., 2005). It consists of 7466 measurements of the abundance of phosphoproteins and phospholipids recorded under different experimental conditions in primary human immune system cells. After processing, 5846 measurements remain. The number of samples under each intervention is shown in Table 1. And the causal structure is presented in Fig. 2.

Table 1. The number of samples under each intervention.

Intervention targets	None	Akt	PKC	PIP2	Mek	PIP3
#samples	1755	911	723	810	799	848

Figure 2. The ground-truth causal graph of the protein dataset.



References

Andersson, S. A., Madigan, D., Perlman, M. D., et al. A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

Eberhardt, F. Causation and intervention. *Unpublished doctoral dissertation, Carnegie Mellon University*, pp. 93, 2007.

He, Y.-B. and Geng, Z. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9:2523–2547, 2008.

Lauritzen, S. L. and Richardson, T. S. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):321–348, 2002.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.