Efficiently Learning and Sampling Intervenional Distributions from Observations

Arnab Bhattacharyya†
arnabb@nus.edu.sg

Sutanu Gayen‡
sutanugayen@gmail.com

Saravanan Kandasamy§
sk3277@cornell.edu

Ashwin Maran¶
amaran@wisc.edu

N. V. Vinodchandran∥
vinod@cse.unl.edu

February 12, 2020

Abstract

We study the problem of efficiently estimating the effect of an intervention on a single variable (atomic interventions) using observational samples in a causal Bayesian network. Our goal is to give algorithms that are efficient in both time and sample complexity in a non-parametric setting.

Tian and Pearl (AAAI ’02) have exactly characterized the class of causal graphs for which causal effects of atomic interventions can be identified from observational data. We make their result quantitative. Suppose $P$ is a causal model on a set $V$ of $n$ observable variables with respect to a given causal graph $G$ with observable distribution $P$. Let $P_x$ denote the interventional distribution over the observables with respect to an intervention of a designated variable $X$ with $x$. We show that assuming that $G$ has bounded in-degree, bounded c-components, and that the observational distribution is identifiable and satisfies certain strong positivity condition:

(i) [Evaluation] There is an algorithm that outputs with probability $2/3$ an evaluator for a distribution $P'$ that satisfies $d_T(P_x, P') \leq \varepsilon$ using $m = \tilde{O}(n\varepsilon^{-2})$ samples from $P$ and $O(mn)$ time. The evaluator can return in $O(n)$ time the probability $P'(v)$ for any assignment $v$ to $V$.

(ii) [Generation] There is an algorithm that outputs with probability $2/3$ a sampler for a distribution $P$ that satisfies $d_T(P_x, P) \leq \varepsilon$ using $m = O(n\varepsilon^{-2})$ samples from $P$ and $O(mn)$ time. The sampler returns an iid sample from $P$ with probability $1 - \delta$ in $O(n\varepsilon^{-1} \log \delta^{-1})$ time.

We extend our techniques to estimate marginals $P_x|_Y$ over a given subset $Y \subseteq V$ of variables of interest. We also show lower bounds for the sample complexity showing that our sample complexity has optimal dependence on the parameters $n$ and $\varepsilon$ as well as the strong positivity parameter.

---

*Author names are in the alphabetical order.
†National University of Singapore. Supported in part by Start-up Grant WBS R252000A33133.
‡National University of Singapore. Supported in part by AB’s Start-up Grant WBS R252000A33133.
§Cornell University. Supported by Cornell University Grant.
¶University of Wisconsin-Madison.
∥University of Nebraska-Lincoln. Research mostly conducted while visiting National University of Singapore.

$P(V \mid do(x))$ is another notation for $P_x$ that is widely used in the literature, with $do(x)$ denoting an intervention on a variable $X$ with value $x$. 

1
1 Introduction

A causal model for a system of variables describes not only how the variables are associated with each other but also how they would change if they were to be acted on by an external force. For example, in order to have a proper discussion about global warming, we need more than just an associational model which would give the correlation between human CO$_2$ emissions and Arctic temperature levels. We instead need a causal model which would predict the climatological effects of humans reducing CO$_2$ emissions by (say) 20% over the next five years. Notice how the two can give starkly different pictures: if global warming is being propelled by natural weather cycles, then changing human emissions won’t make any difference to temperature levels, even though human emissions and temperature may be correlated in our dataset (just because both are increasing over the timespan of our data).

Causality has been a topic of inquiry since ancient times, but a modern, rigorous formulation of causality came about in the twentieth century through the works of Pearl, Robins, Rubin, and others [IR15, Pea09, RSR11, HR20]. In particular, Pearl [Pea09] recasted causality in the language of causal Bayesian networks (or causal Bayes nets for short). A causal Bayes net is a standard Bayes net that is reinterpreted causally. Specifically, it makes the assumption of modularity: for any variable $X$, the dependence of $X$ on its parents is an autonomous mechanism that does not change even if other parts of the network are changed. This allows assessment of external interventions, such as those encountered in policy analysis, treatment management, and planning. The idea is that by virtue of the modularity assumption, an intervention simply amounts to a modified Bayes net where some of the parent-child mechanisms are altered while the rest are kept the same.

The underlying structure of causal Bayes net $\mathcal{P}$ is a directed acyclic graph $G$. The graph $G$ consists of $n+h$ nodes where $n$ nodes correspond to the observable variables $V$ while the $h$ additional nodes correspond to a set of $h$ hidden variables $U$. We assume that the observable variables take values over a finite alphabet $\Sigma$. By interpreting $\mathcal{P}$ as a standard Bayes net over $V \cup U$ and then marginalizing to $V$, we get the observational distribution $P$ on $V$. The modularity assumption allows us to define the result of an intervention on $\mathcal{P}$. An intervention is specified by a subset $X \subseteq V$ of variables and an assignment $x \in \Sigma^{|X|}$. In the interventional distribution, the variables $X$ are fixed to $x$, while each variable $W \in (V \cup U) \setminus X$ is sampled as it would have been in the original Bayes net, according to the conditional distribution $W \mid \text{Pa}(W)$, where $\text{Pa}(W)$ (parents of $W$) consist of either variables previously sampled in the topological order of $G$ or variables in $X$ set by the intervention. The marginal of the resulting distribution to $V$ is the interventional distribution denoted by $P_x$. We sometimes also use $\text{do}(x)$ to denote the intervention process and $P(V \mid \text{do}(x))$ to denote the resulting interventional distribution.

In this work, we focus our attention on the case that $X$ is a single observable variable, so that interventions on $X$ are atomic. We study the following estimation problems:

1. **(Evaluation)** Given an $x \in \Sigma$, construct an evaluator for $P_x$ which estimates the value of the probability mass function

$$P_x(v) \overset{\text{def}}{=} \Pr_{V \sim P_x} [V = v]$$

for any $v \in \Sigma^n$. The goal is to construct the evaluator using only a bounded number of samples from the observational distribution $P$, and moreover, the evaluator should run efficiently.

---

$^2$Consistent with the convention in the causality literature, we will use a lower case letter (e.g., $x$) to denote an assignment to the subset of variables corresponding to its upper case counterpart (e.g., $X$).
2. **(Generation)** Given an \( x \in \Sigma \), construct a *generator* for \( P_x \) which generates i.i.d. samples from a distribution that approximates \( P_x \). The goal is to construct the generator using only a bounded number of samples from the observational distribution \( P \), and moreover, the generator should be able to output each sample efficiently.

We study these problems in the non-parametric setting, where we assume that all the observable variables under consideration are over a finite alphabet \( \Sigma \).

Evaluation and generation are two very natural inference problems\(^3\). Indeed, the influential work of Kearns et al. [KMR+94] introduced the computational framework of *distribution learning* in terms of these two problems. Over the last 25 years, work on distribution learning has clarified how classical techniques in statistics can be married to new algorithmic ideas in order to yield sample- and time-efficient algorithms for learning very general classes of distributions; see [Dia16] for a recent survey of the area. The goal of our work is to initiate a similar computational study of the fundamental problems in causal inference.

The crucial distinction of our setting from the distribution learning setting is that the algorithm does not get samples from the distribution of interest. In our setting, the algorithm receives as input samples from \( P \) while its goal is to estimate the distribution \( P_x \). This is motivated by the fact that typically randomized experiments are hard (or unethical) to conduct while observational samples are easy to collect. Even if we disregard computational considerations, it may be impossible to determine the interventional distribution \( P_x \) from the observational distribution \( P \) and knowledge of the causal graph \( G \). The simplest example is the so-called “bow-tie graph” on two observable variables \( X \) and \( Y \) (with \( X \) being a parent of \( Y \)) and a hidden variable \( U \) that is a parent of both \( X \) and \( Y \). Here, it’s easy to see that \( P \) does not uniquely determine \( P_x \). Tian and Pearl [TP02a] studied the general question of when the interventional distribution \( P_x \) is identifiable from the observational distribution \( P \). They characterized the class \( G_X \) of directed acyclic graphs with hidden variables such that for any \( G \in G_X \), for any causal Bayes net \( P \) on \( G \), and for any intervention \( x \) to \( X \), \( P_x \) is identifiable from \( P \). Thus, for all our work we assume that \( G \in G_X \), because otherwise, \( P_x \) is not identifiable, even with an infinite number of observations.

We design sample and time efficient algorithms for the above-mentioned estimation problems. Our starting point is the work of Tian and Pearl [TP02a]. Tian and Pearl [TP02a] (as well as other related work on identifiability) assumes, in addition to the underlying graph being in \( G_X \), that the distribution \( P \) is *positive*, meaning that \( P(\mathbf{v}) > 0 \) for all assignments \( \mathbf{v} \) to \( \mathbf{V} \). We show that under reasonable assumptions about the structure of \( G \), we only need to assume *strong positivity* for the marginal of \( P \) over a bounded number of variables to design our algorithms. We extend our techniques to the problem of efficiently estimating the marginal interventional distributions over a subset of observable variables. Finally we establish a lower bound for the sample complexity showing that our sample complexity has near optimal dependence on the parameters of interest. We discuss our results in detail next.

## 2 Our Contributions

Let \( \mathcal{P} \) be a causal Bayes net\(^4\) over a graph \( G \), in which the set of observable variables is denoted by \( \mathbf{V} \) and the set of hidden variables is denoted by \( \mathbf{U} \). Let \( n = |\mathbf{V}| \). There is a standard procedure in

\(^3\)Note that the distinction between the two problems is computational; one can produce a generator from an evaluator and vice versa without requiring any new samples.

\(^4\)Formal definitions appear in Section 3.
the causality literature (see [TP02b]) to convert $G$ into a graph on $n$ nodes. Namely, under the semi-Markovian assumption that each hidden variable $U$ does not have any parents and affects exactly two observable variables $X_i$ and $X_j$, we remove $U$ from $G$ and put a bidirected edge between $X_i$ and $X_j$. We end up with an Acyclic Directed Mixed Graph (ADMG) $G$, having $n$ nodes corresponding to the variables $V$ and having edge set $E\rightarrow \cup E\leftrightarrow$ where $E\rightarrow$ are the directed edges and $E\leftrightarrow$ are the bidirected edges. Figure 1 shows an example. The in-degree of $G$ is the maximum number of directed edges coming into any node. A c-component refers to any maximal subset of nodes/variables which is connected using only bidirected edges. Then $V$ gets partitioned into c-components: $S_1, S_2, \ldots, S_\ell$.

Let $X$ be a designated variable in $V$. Without loss of generality, suppose $X \in S_1$. Tian and Pearl [TP02a] showed that $G_X$ (the class of ADMGs $G$ for which $P_x$ is identifiable from $P$ for any causal Bayes net $P$ on $G$ and any intervention $x \in \Sigma$) consists of exactly those graphs that satisfy Assumption 2.1 below (See Theorem 3 of [TP02a]).

The second assumption we make is about the observational distribution $P$. For a subset of variables $S \subseteq V$, let $\text{Pa}^+(S) = S \cup \bigcup_{V \in S} \text{Pa}(V)$ where $\text{Pa}(V)$ are the observable parents of $V$ in the graph $G$.

**Assumption 2.2** ($\alpha$-strong positivity with respect to $X$). Suppose $X$ lies in the c-component $S_1$, and let $Z = \text{Pa}^+(S_1)$. For every assignment $z$ to $Z$, $P(Z = z) > \alpha$.

So, if $|\text{Pa}^+(S_1)|$ is small, then Assumption 2.2 only requires that a small set of variables take on each possible configuration with non-negligible probability. When Assumption 2.2 holds, we say that the causal Bayes net is $\alpha$-strongly positive with respect to $X$.

### 2.1 Algorithms

Suppose $P$ is an unknown causal Bayes net over a known ADMG $G$ on $n$ observable variables $V$ that satisfies identifiability (Assumption 2.1) and $\alpha$-strong positivity (Assumption 2.2) with respect to a variable $X \in V$. Let $d$ denote the maximum in-degree of the graph $G$ and $k$ denote the size of its largest c-component.

We first present an efficient algorithm for the evaluation problem.

**Theorem 2.3** (Evaluation). For any intervention $x$ to $X$ and parameter $\varepsilon \in (0,1)$, there is an algorithm that takes $m = \tilde{O}\left(\frac{|\Sigma|^{kd}n}{\alpha\varepsilon^2}\right)$ samples from $P$, and in $O(mn|\Sigma|^{2kd})$ time, returns a circuit
For any subset $F \subseteq V$ with $|F| = f$, intervention $x$ to $X$ and parameter $\varepsilon \in (0,1)$, there is an algorithm that takes $m = \tilde{O}\left(\frac{|\Sigma|^k n^{f+k(d+1)^2}}{\alpha \varepsilon^2}\right)$ samples from $P$ and returns an evaluator for a distribution $\hat{P}_F$ on $F$ such that $d_{TV}(\hat{P}_F, \hat{P}_x|F) \leq \varepsilon$.

Note that the time complexity of the above algorithm is exponential in $f$ as we need to take exponential in $f$ many samples from the generator. To handle problems that arise in practice for $F$'s of small cardinality, it is of interest to develop sample and time efficient algorithms for estimating $P_x|F$. In such cases the approach discussed above is superfluous, as the sample complexity depends linearly on $n$, the total number of variables in the model, which could be potentially large. We show that in such cases where $f$ is extremely small we can perform efficient estimation with small sample size. A more detailed discussion of our analysis on evaluating marginals which includes the algorithms and proofs can be found in Section 6. Precisely, we show the following theorem:

**Theorem 2.6.** For any subset $F \subseteq V$ with $|F| = f$, intervention $x$ to $X$ and parameter $\varepsilon \in (0,1)$, there is an algorithm that takes $m = \tilde{O}\left(\frac{|\Sigma|^k n^{f+k(d+1)^2}}{\alpha \varepsilon^2}\right)$ samples from $P$ and runs in $O(m(f + k(d + 1))|\Sigma|^2(f+k(d+1)^2)$ time and returns an evaluator for a distribution $\hat{P}_F$ on $F$ such that $d_{TV}(\hat{P}_F, \hat{P}_V) \leq \varepsilon$.

### 2.2 Lower Bounds

We next address the question of whether the sample complexity of our algorithms has the right dependence on the parameters of the causal Bayes net as well as on $\alpha$. We also explore whether
that \( \Omega(\epsilon^{-2}) \) samples \((X,Y)\) with \( X = 0 \) are needed. Since \( X = 0 \) occurs with probability \( \alpha \), an \( \Omega(\alpha^{-1}\epsilon^{-2}) \) lower bound on the sample complexity follows.

However, from this example, it’s not clear that we need to enforce strong positivity on the parents of \( X \) or the \( c \)-component containing \( X \), since both are trivial. Also, the sample complexity has no dependence on \( n \) and \( d \). The following theorem addresses these issues.

**Theorem 2.7.** Fix integers \( d, k \geq 1 \) and a set \( \Sigma \) of size \( \geq 2 \). For all sufficiently large \( n \), there exists an ADMG \( G \) with \( n \) nodes and in-degree \( d \) so that the following hold. \( G \) contains a node \( X \) such that \( |\text{Pa}(X)| = d \) and \( |S_1| = k \) (where \( S_1 \) is the \( c \)-component containing \( X \)). For any \( Z \in \text{Pa}(X) \cup S_1 \), there exists a causal Bayes net \( P \) on \( G \) over \( \Sigma \)-valued variables such that:

1. For the observational distribution \( P \), the marginal \( P|_{\text{Pa}(X) \cup S_1 \setminus \{Z\}} \) is uniform but the marginal \( P|_{\text{Pa}(X) \cup S_1} \) has mass at most \( \alpha \) at some assignment.
2. There exists an intervention \( x \) on \( X \) such that learning the distribution \( P_x \) up to \( d_{TV} \)-distance \( \epsilon \) with probability \( 9/10 \) requires \( \Omega(n|\Sigma|^d/\alpha\epsilon^2) \) samples from \( P \).

So, \( P \) must have a guarantee that its marginal on \( \text{Pa}(X) \cup S_1 \) has mass \( > \alpha \) at all points in order for an algorithm to learn \( P_x \) using \( O(n|\Sigma|^d/\alpha\epsilon^2) \) samples. For comparison, our algorithms in Theorem 2.3 and Theorem 2.4 assume strong positivity for \( \text{Pa}^+(S_1) \) and achieve sample complexity \( O(n|\Sigma|^5d/\alpha^2) \). For small values of \( k \) and \( d \), the upper and lower bounds are qualitatively similar. It remains an open question to fully close the gap.

To hint towards the proof of Theorem 2.7, we sketch the argument when \( Z \) is a parent of \( X \) and \( n = 3 \). Figure 2b shows a graph where \( X \) has one parent \( Z \) and no hidden variables. Both \( X \) and \( Z \) are parents of \( Y \), and all three are binary variables. Consider two causal models \( P \) and \( Q \). For both \( P \) and \( Q \), \( Z \) is uniform over \( \{0,1\} \) and \( X \neq Z \) with probability \( \alpha \). Now, suppose \( P(Y = 1 \mid X \neq Z) = 1/2 + \epsilon \) and \( P(Y = 1 \mid X = Z) = 1/2 \), while \( Q(Y = 1 \mid X \neq Z) = 1/2 - \epsilon \) and \( Q(Y = 1 \mid X = Z) = 1/2 \). Note that \( P_{X=1}(Y = 1) = (1 + \epsilon)/2 \) while \( Q_{X=1}(Y = 1) = (1 - \epsilon)/2 \), so that the interventional distributions are \( \epsilon \)-far from each other. On the other hand, it can be shown using Fano’s inequality that any algorithm needs to observe \( \Omega(\alpha^{-1}\epsilon^{-2}) \) samples to distinguish \( P \) and \( Q \).
2.3 Previous Work

Identification of causal effects from the observational distribution has been studied extensively in the literature. Here we discuss some of the relevant literature in the non-parametric setting. When there are no unobservable variables (and hence the associated ADMG is a DAG), it is always possible to identify any given intervention from the observational distribution [Pea09, Rob86, SGS00]. However, when there are unobservable variables causal effect identifiability in ADMGs is not always possible. A series of important works focused on establishing graphical criterions for identifiability of interventional distributions from the observational distribution [TP02a, SGS00, GP95, Hal00, KM99, PR95]. This led to a complete algorithm, first by Tian and Pearl for the identifiability of atomic interventions [TP02a] (this work is the most relevant for the present work), and then by Shpitser and Pearl (algorithm ID) for the identifiability of any given intervention from the observational distribution [SP06] (see also [HV08]). Researchers have also investigated implementation aspects of the identification algorithms. In particular, an implementation of the algorithm ID has been carried out in the R package causaleffect in [TK17a]. This work was followed by a sequence of works [TK17b, TK18] where the authors simplify ID and obtain a succinct representation of the target causal effect by removing unnecessary variables from the expression. Other software packages related to causal identifiability are also publicly available [Tia, Kel, SK].

Researchers have also investigated non-parametric causal effect identification from observations on structures other than ADMGs. Some recent results in this direction include work reported in [JZB19a] (and [JZB19b]) where complete algorithms have been established for causal effect identifiability (and conditional causal effect identifiability) with respect to Markov equivalent class diagrams, a more general class of causal graphs. Maximally oriented partially directed acyclic graphs (MPDAGs) is yet another generalization of DAGs with no hidden variables. Very recently complete algorithms for causal identification with respect to MPDAGs have been established [Per19]. Complete algorithms are also known for dynamic causal networks, a causal analogue for dynamic Bayesian networks that evolve over time [BAG16]. Causal chain graphs (CEGs, which are similar to ADMGs) are yet another class of graphs for which identifiability of interventions has been investigated and conditions (similar to Pearl’s back-door criterion) have been established [TSR10, Thw13].

In a different line of work reported in [SS16], the authors introduce the notion of stability of causal identification: a notion capturing the sensitivity of causal effects to small perturbations in the input. They show that the causal identification function is numerically unstable for the ID algorithm [SP06]. They also show that, in contrast for atomic interventions (i.e., when \(X\) is singleton) the identification algorithm of Tian and Pearl [TP02a] is not too sensitive to changes in the input whenever Assumption 2.1 of [TP02a] is true.

Although most of the work on non-parametric causal identification mentioned above assume the causal graph is known, the problem of inferring the underlying causal graph has also been studied in various contexts. Some papers reporting the work along this line include [HEJ15, HIB13, ASY+19, YKU18, KJSB19]. Causal effect identification is a fundamental topic with a wide range of practical applications. In particular it has found applications in a range of applied areas including recommendation systems [SHW15], computational sciences [Spi10], social and behavioral sciences [Sob00], econometrics [HV07, Mat93, Lew19], and epidemiology [HR20].

An important observation we note is that all existing works on non-parametric causal identifiability research assume infinite sample access to the observational distribution. To the best of our

---

5Complete algorithms output the desired causal effect whenever possible or output fail along with a proof of unidentifiability – thus characterizing a necessary and sufficient graphical condition for identifiability.
knowledge, the present work is the first that establishes sample and time complexity bounds on non-parametric causal effect identifiability. In this respect, the closest related work is [ABDK18] which looked at the problem of goodness-of-fit testing of causal models in a non-parametric setting; however, they assumed access to experimental data, not just observational data.

3 Preliminaries

Notation. We use capital (bold capital) letters to denote variables (sets of variables), e.g., $A$ is a variable and $B$ is a set of variables. We use small (bold small) letters to denote values taken by the corresponding variables (sets of variables), e.g., $a$ is the value of $A$ and $b$ is the value of the set of variables $B$. For a vector $v$ and a subset of coordinates $S$, we use the notation $v_S$ to denote the restriction of $v$ to the coordinates in $S$ and $v_i$ to denote the $i$-th coordinate of $v$. For two sets of variables $A$ and $B$ and assignments of values $a$ to $A$ and $b$ to $B$, $a \circ b$ denotes the assignment to $A \cup B$ in the natural way.

The variables in this paper take values in a finite set $\Sigma$. We use the total variation distance to measure the distances between distributions. For two distributions $P$ and $Q$ over the same finite sample space $[D]$, their total variation distance is denoted by $d_{TV}(P, Q)$ and is given by $d_{TV}(P, Q) := \frac{1}{2} \sum_{i \in [D]} |P(i) - Q(i)|$.

Bayesian Networks. Bayesian networks are popular probabilistic graphical models for describing high-dimensional distributions.

Definition 3.1. A Bayesian Network $P$ is a distribution that can be specified by a tuple $\langle V, G, \{\Pr[V_i | \pa(V_i)] : V_i \in V, \pa(V_i) \in \Sigma^{\pa(V_i)}\}\rangle$ where: (i) $V = (V_1, \ldots, V_n)$ is a set of variables over alphabet $\Sigma$, (ii) $G$ is a directed acyclic graph with $n$ nodes corresponding to the elements of $V$, and (iii) $\Pr[V_i | \pa(V_i)]$ is the conditional distribution of variable $V_i$ given that its parents $\pa(V_i)$ in $G$ take the values $\pa(V_i)$.

The Bayesian Network $P = \langle V, G, \{\Pr[V_i | \pa(V_i)]\}\rangle$ defines a probability distribution over $\Sigma^{|V|}$, as follows. For all $v \in \Sigma^{|V|}$,

$$P(v) = \prod_{V_i \in V} \Pr[v_i | \pa(V_i) = v_{\pa(V_i)}].$$

In this distribution, each variable $V_i$ is independent of its non-descendants given its parents in $G$.

Causality. We describe Pearl’s notion of causality from [Pea95]. Central to his formalism is the notion of an intervention. Given a variable set $V$ and a subset $X \subset V$, an intervention $\text{do}(x)$ is the process of fixing the set of variables $X$ to the values $x$. The interventional distribution $P_x$ is the distribution on $V$ after setting $X$ to $x$. Formally:

Definition 3.2 (Causal Bayes Net). A causal Bayes net $\mathcal{P}$ is a collection of interventional distributions that can be defined in terms of a tuple $\langle V, U, G, \{\Pr[V_i | \pi(V_i)] : V_i \in V, \pi(V_i) \in \Sigma^{|\Pi(V_i)|}\}, \{\Pr[U]\}\rangle$, where (i) $V = (V_1, \ldots, V_n)$ and $U$ are the tuples of observable and hidden variables respectively, (ii) $G$ is a directed acyclic graph on $V \cup U$, (iii) $\Pr[V_i | \pi(V_i)]$ is the conditional probability distributions of $V_i \in V$ given that its parents $\Pi(V_i) \in V \cup U$ take the values $\pi(V_i)$, and (iv) $\Pr[U]$ is the distribution of the hidden variables $U$. $G$ is said to be the causal graph corresponding to $\mathcal{P}$.
Such a causal Bayes net $\mathcal{P}$ defines a unique interventional distribution $P_x$ for every subset $X \subseteq V$ (including $X = \emptyset$) and assignment $x \in \Sigma^{|X|}$, as follows. For all $v \in \Sigma^{|V|}$:

$$P_x(v) = \begin{cases} 
\sum_u \prod_{v_i \in V \setminus X} \Pr[v_i \mid \Pi(V_i) = v_{\Pi(V_i)}] \cdot \Pr[u] & \text{if } v \text{ is consistent with } x \\
0 & \text{otherwise.}
\end{cases}$$

We use $P$ to denote the observational distribution ($X = \emptyset$). For a subset $F \subseteq V$, $P_x|_F$ denotes the marginal of $P_x$ on $F$. For an assignment $f$ to $F$, we also use the notation $P_x(f)$ as shorthand for the probability mass of $P_x|_F$ at $f$.

As mentioned in the introduction, we often consider a causal graph $G$ as an ADMG by implicitly representing hidden variables using bidirected edges. In an ADMG, we imagine that there is a hidden variable subdividing each such bidirected edge that is a parent of the two endpoints of the edge. Thus, the edge set of an ADMG is the union of the directed edges $E^\rightarrow$ and the bidirected edges $E^{++}$. Given such an ADMG $G$, for any $S \subseteq V$, $\overline{S}$ denotes the complement set $V \setminus S$, $\text{Pa}(S)$ denotes the parents of $S$ according to the directed edges of $G$, i.e., $\text{Pa}(S) = \cup_{X \in S} \{Y \in V : (Y, X) \in E^\rightarrow\}$. We also define: $\text{Pa}^+(S) = \text{Pa}(S) \cup S$ and $\text{Pa}^-(S) = \text{Pa}(S) \setminus S$. The bidirected edges are used to define c-components:

**Definition 3.3 (c-component).** For a given ADMG $G$, $S \subseteq V$ is a c-component of $G$, if $S$ is a maximal set such that between any two vertices of $S$, there exists a path that uses only the bidirected edges $E^{++}$.

Since a c-component forms an equivalence relation, the set of all c-components forms a partition of $V$, the observable vertices of $G$. Let $S_1 \cup S_2 \cup \cdots \cup S_\ell$ denote the partition of $V$ into the c-components of $G$.

**Definition 3.4.** For a subset $S \subseteq V$, the Q-factor for $S$ is defined as the following function over $\Sigma^{|V|}$:

$$Q_S(v) = P_{v_S}(v_S).$$

Clearly, for every $v_S$, $Q_S$ is a distribution over $\Sigma^{|S|}$.

For $Y \subseteq V$, the induced subgraph $G|_Y$ is the subgraph obtained by removing the vertices $V \setminus Y$ and their corresponding edges from $G$.

The following lemma is used heavily in this work.

**Lemma 3.5** (Corollary 1 of [Tia02]). Let $\mathcal{P}$ be a causal Bayes net on $G = (V, E^\rightarrow \cup E^{++})$. Let $S_1, \ldots, S_\ell$ be the c-components of $G$. Then for any $v$ we have:

(i) $P(v) = \prod_{j=1}^\ell Q_{S_j}(v)$.

(ii) Let $V_1, V_2, \ldots, V_\ell$ be a topological order over $V$ with respect to the directed edges. Then, for any $j \in [\ell]$, $Q_{S_j}(v)$ is computable from $P(v)$ and is given by:

$$Q_{S_j}(v) = \prod_{i : v_i \in S_j} P(v_i \mid v_1, \ldots, v_{i-1}).$$
(iii) Furthermore, each factor $P(v_i \mid v_1, \ldots, v_{i-1})$ can be expressed as:

$$P(v_i \mid v_1, \ldots, v_{i-1}) = P(v_i \mid v_{\text{Pa}(T_i) \cap \{i-1\}})$$

where $T_i$ is the c-component of $G[V_1, \ldots, V_i]$ that contains $V_i$.

Note that Lemma 3.5 implies that each $Q_S_j(v)$ is a function of the coordinates of $v$ corresponding to $\text{Pa}^+(S_j)$. The next result, due to Tian and Pearl, uses the identifiability criterion encoded in Assumption 2.1.

**Theorem 3.6** (Theorem 3 of [TP02a]). Let $P$ be a causal Bayes net over $G = (V, E^\rightarrow \cup E^{\leftrightarrow})$ and $X \in V$ be a variable. Let $S_1, \ldots, S_\ell$ be the c-components of $G$ and assume $X \in S_1$ without loss of generality. Suppose $G$ satisfies Assumption 2.1 (identifiability with respect to $X$). Then for any setting $x$ to $X$ and any assignment $w$ to $\{X\}^c$, the interventional distribution $P_x(w)$ is given by:

$$P_x(w) = \sum_{x' \in \Sigma} Q_{S_1}(w \circ x') \cdot \prod_{j=2}^\ell Q_{S_j}(w \circ x)$$

# 4 Efficient Estimation

Let $P$ be a causal Bayes net over a causal graph $G = (V, E^\rightarrow \cup E^{\leftrightarrow})$. $G$ is an ADMG with observable variables $V_1, \ldots, V_n$. Without loss of generality, let $V_1, \ldots, V_n$ be a topological order according to the directed edges of $G$. As a first step to our algorithms for interventional distributions, we are interested in learning the observational distribution $P(V)$. Our approach is to view the causal Bayes net as a regular Bayes net over observable variables and use an existing learning algorithm for Bayes nets. From Lemma 3.5, we can write the observational distribution $P(V)$ as:

$$P(V) = \prod_{i=1}^nP(V_i \mid Z_i)$$

where $Z_i \subseteq \{V_1, \ldots, V_{i-1}\}$ is of size at most $kd+k$. Here $k$ is the maximum c-component size and $d$ is the maximum in-degree. Therefore $P$ can also be viewed as the distribution of a (regular) Bayes net with no hidden variables but with in-degree at most $kd+k$. The problem of properly learning a Bayes net is well-studied [Das97, CDKS17], starting from Dasgupta’s early work [Das97]. We use the following learning result described in [BGMV20].

**Theorem 4.1** ([BGMV20]). There is an algorithm that on input parameters $\varepsilon, \delta \in (0, 1)$ and samples from an unknown Bayes net $P$ over $\Sigma^n$ on a known DAG $G$ on vertex set $[n]$ and maximum in-degree $d$, takes $m = O(\log \frac{1}{\delta} |\Sigma|^{d+1}n \log(n|\Sigma|^{d+1})/\varepsilon^2)$ samples, runs in time $O(mn|\Sigma|^{d+1})$, and produces a Bayes net $\hat{P}$ on $G$ such that $d_{TV}(P, \hat{P}) \leq \varepsilon$ with probability $\geq 1 - \delta$.

From the above discussion we get the following corollary.
Corollary 4.2. There is an algorithm that on input parameters \( \epsilon, \delta \in (0, 1) \), and samples from the observed distribution \( P \) of an unknown causal Bayes net over \( \Sigma^n \) on a known ADMG \( G \) on vertex set \([n]\) with maximum in-degree \( d \) and maximum c-component size \( k \), takes \( m = \tilde{O}(\frac{k}{\epsilon^2} |\Sigma|^{kd+k+1} \log \frac{1}{\delta}) \) samples, runs in time \( O(mn|\Sigma|^{kd+k+1}) \) and outputs a Bayes net \( \hat{P} \) on a DAG \( G' \) such that \( d_{TV}(P, \hat{P}) \leq \epsilon \) with probability \( \geq 1 - \delta \).

Lemma 3.5 together with the Bayes net learning algorithm also gives us a way to learn the joint distribution on a subset of c-components when the remaining observable variables are intervened (assuming certain strong positivity condition), which we will use.

Lemma 4.3. Let \( S \subseteq \{S_1, \ldots, S_t\} \) be any fixed set of c-components of an ADMG \( G \) on vertex set \([n]\) with maximum in-degree \( d \) and maximum c-component size \( k \). Let \( Y \) be the set of vertices in \( S \) and \( \overline{Y} = V \setminus Y \). Let \( \overline{Y} \) be any assignment of \( Y \) and \( \overline{Y}_{Pa} \) be the restriction of \( \overline{Y} \). Suppose \( P(\overline{Y}_{Pa}) \geq \alpha \). Then, (1) \( P(Y) = P_{Y_{Pa}}(Y) \). (2) Moreover, there is an algorithm that on input parameters \( \epsilon, \delta \in (0, 1) \), and samples from the observed distribution \( P \) on \( G \) takes \( m = \tilde{O}(\log \frac{1}{\delta} |\Sigma|^{kd+k+1}|Y|/\alpha \epsilon^2) \) samples, runs in time \( O(m|Y||\Sigma|^{kd+k+1}) \) and outputs a Bayes net \( \hat{P} \) on a DAG \( G' \) such that \( d_{TV}(\hat{P}(Y), \hat{P}) \leq \epsilon \) with probability \( \geq 1 - \delta \).

Proof. \( P_Y(Y) = P_{Y_{Pa}}(Y) \) follows from the fact that the intervention \( Y_{Pa} \) blocks any other interventions since these are the immediate parents of \( Y \) in \( G \). An informal argument to show \( P_Y(Y) \) is identifiable implicitly follows from Lemma 3.5. We can put dummy bidirected edges to make \( S \) a single c-component and apply part (ii) of Lemma 3.5. In the resulting factorization the terms are of the form \( P(v_i | v_{i-1}) \). But then we can only keep the subset of variables of \( v_{i-1} \) which are from the c-component of \( v_i \) and the parents of this c-component, and remove the rest using conditional independence (blocking), since the dummy bidirected edges are not affecting.

Hence for a fixed assignment \( Y_{Pa} \), the factorization of the joint distribution \( P_Y(Y) \) consists of the product of only the terms \( P(V_i | Z_i) \) for every \( i \in Y \), where the set \( Z_i \) has at most \( kd + k \) variables among which the variables coming from \( Pa^{-1}(Y) \) are fixed to \( Y_{Pa} \). More formally this fact can be proven by closely following the proof of Lemma 3.5.

It follows by comparing the above factorization of \( P_Y(Y) \) and the factorization of \( P(V) \) from Lemma 3.5 that the conditional samples from \( P(V) \) with \( Pa^{-1}(Y) = Y_{Pa} \) is according to the distribution \( P_Y(Y) \). Moreover \( P_Y(Y) \) is a Bayes net with degree at most \( kd + k \) as we discussed in the previous paragraph. From Theorem 4.1 \( m' = O(\log \frac{1}{\delta} |\Sigma|^{kd+k+1}|Y| \log(|Y||\Sigma|^{kd+k+1})/\epsilon^2) \) conditional samples and processing time \( O(m'|Y||\Sigma|^{kd+k+1}) \) is enough to learn in \( d_{TV} \) distance \( \leq \epsilon \). The result follows by observing \( O(m'/\alpha) \) samples from \( P \) are enough to give \( m \) conditional samples except for probability \( \leq \delta \).

In the next two subsections we design our evaluation and generation algorithms. We use the following partitioning of the variables in these subsections. Let the c-components of \( G \) be given by \( S_1, \ldots, S_t \) with \( X \in S_1 \). Let \( A = S_1 \setminus \{X\} \), \( B = Pa^{-1}(S_1) \) and \( C = V \setminus (A \cup B \cup \{X\}) \). Note that \(|A| \leq k \) and \(|B| \leq kd \). See Figure 3.
4.1 Evaluation

Algorithm 1: Algorithm to output evaluator for $P_x$. For notations, please refer to the proof of Theorem 2.3 in Section 4.1

**Input**: Sample access to $P$

**Output**: Evaluator for $P_x$

1. for every assignment $a$ of $A = S_1 \setminus \{X\}$ do
2. \[ S \leftarrow \text{Draw } \Theta(n|\Sigma|^{kd+3k+1}/\alpha\varepsilon^2) \text{ samples from } P; \]
3. if less than $\Theta(n|\Sigma|^{kd+3k+1}/\varepsilon^2)$ samples with $s_1 = a \circ x$ appear in $S$ then
4. \[ \text{FAIL;} \]
5. else
6. \[ R'_a(B, C) \leftarrow \text{Learn } P_{s_1}(B \circ C) \text{ using Lemma 4.3;} \]

7. for every assignment $b$ to $B = Pa^{-}(S_1)$ do
8. \[ S \leftarrow \text{Draw } \Theta(k|\Sigma|^{kd+3k+1}/\alpha\varepsilon^2) \text{ samples from } P; \]
9. if less than $\Theta(k|\Sigma|^{kd+3k+1}/\varepsilon^2)$ samples with $b$ appear in $S$ then
10. \[ \text{FAIL;} \]
11. else
12. \[ M'_b(A) \leftarrow \text{Learn } P_b(A) \text{ using Lemma 4.3;} \]

13. return a function that, on input an assignment $w = a \circ b \circ c$ to $W = V \setminus \{X\}$, outputs $M'_b(a) \cdot R'_a(b, c)$;

**Proof of Theorem 2.3.** In view of Theorem 3.6, for an assignment $w = a \circ b \circ c$ ($w_A = a$, $w_B = b$, $w_C = c$) to $V \setminus \{x\}$ let us define $M$ and $R$ so that $P_x(w) = M(w)R(w)$:

\[
M(w) = P_{w_B}(w_A) = \sum_{x'} P_{w_B}(w_A \circ x') = \sum_{x' \in \Sigma} Q_{S_1}(w \circ x')
\]

\[
R(w) = P_{w_A \circ x}(w_{B \cup C}) = \prod_{j=2}^{\ell} Q_{S_j}(w \circ x).
\]

Our strategy will be to approximate $R$ and $M$ separately as $R'$ and $M'$, and then output their product. See Algorithm 1.

For every assignment $a$ to $A$, let $R_a$ denote the partial function of $R$ for those $w$ such that $w_A = a$. Note that we can interpret $R_a$ as a probability distribution on $B \cup C$, thus $R(a, b, c) = R_a(b, c)$. Invoking Lemma 4.3 and using Assumption 2.2, for every $a$, we can learn an $R'_a$ such that $d_{TV}(R'_a, R_a) < \varepsilon/2|\Sigma|^k$, using $\tilde{O}(|\Sigma|^{5kd/n}/\alpha\varepsilon^2)$ samples.
For a particular assignment \( \mathbf{b} \) to \( \mathbf{B} \), let \( M_{\mathbf{b}} \) denote the partial function of \( M \) for those \( \mathbf{w} \) such that \( \mathbf{w}_\mathbf{B} = \mathbf{b} \). We can interpret \( M_{\mathbf{b}} \) as a probability distribution on \( \mathbf{A} \), thus \( M(\mathbf{a}, \mathbf{b}, \mathbf{c}) = M_{\mathbf{b}}(\mathbf{a}) \). We learn the distribution on all of \( \mathbf{S}_1 \) and then take its marginal on \( \mathbf{A} \). Again, we can invoke Lemma 4.3 and Assumption 2.2, so that for every \( \mathbf{b} \), we learn an \( M'_{\mathbf{b}} \) that is within \( \varepsilon/2|\Sigma|^k \) TV distance of \( M_{\mathbf{b}} \).

Finally, we use Lemma 4.4 below to combine the two estimates to obtain an estimate \( P'_x(\mathbf{w}) := M'(\mathbf{w})R'(\mathbf{w}) \).

**Lemma 4.4.** Suppose for all assignments \( \mathbf{b} \) to \( \mathbf{B} \), \( d_{TV}(M_{\mathbf{b}}(\mathbf{A}), M'_{{\mathbf{b}}}({\mathbf{A}})) \leq \varepsilon_1 \) and for all assignments \( \mathbf{a} \) to \( \mathbf{A} \), \( d_{TV}(R_{\mathbf{a}}(\mathbf{B}, \mathbf{C}), R'_{{\mathbf{a}}}({\mathbf{B}}, \mathbf{C})) \leq \varepsilon_2 \). Then \( \sum_{\mathbf{w} \in \Sigma^{n-1}} |P_x(\mathbf{w}) - P'_x(\mathbf{w})| \leq |\Sigma|^{k(\varepsilon_1 + \varepsilon_2)} \).

**Proof.**

\[
\sum_{\mathbf{w}} |P_x(\mathbf{w}) - P'_x(\mathbf{w})| = \sum_{\mathbf{w}} |M(\mathbf{w})R(\mathbf{w}) - M'(\mathbf{w})R'(\mathbf{w})| \\
= \sum_{\mathbf{w}} |M(\mathbf{w})R(\mathbf{w}) - M(\mathbf{w})R'(\mathbf{w}) + M(\mathbf{w})R'(\mathbf{w}) - M'(\mathbf{w})R'(\mathbf{w})| \\
\leq \sum_{\mathbf{w}} |M(\mathbf{w})||R(\mathbf{w}) - R'(\mathbf{w})| + |M(\mathbf{w}) - M'(\mathbf{w})||R'(\mathbf{w})| \\
\leq \sum_{\mathbf{w}} |M(\mathbf{w})||R(\mathbf{w}) - R'(\mathbf{w})| + |M(\mathbf{w}) - M'(\mathbf{w})| \leq \varepsilon/2|\Sigma|^k
\]

We upper bound each of the above two terms separately. We break \( \mathbf{w} = \mathbf{a} \circ \mathbf{b} \circ \mathbf{c} \) corresponding to \( \mathbf{A}, \mathbf{B} \) and \( \mathbf{C} \) respectively.

\[
\sum_{\mathbf{a}, \mathbf{b}, \mathbf{c}} |M_{\mathbf{b}}(\mathbf{a}) - M'_{{\mathbf{b}}}({\mathbf{a}})||R_{\mathbf{a}}(\mathbf{b}, \mathbf{c}) = \sum_{\mathbf{a}, \mathbf{b}, \mathbf{c}} (\sum_{\mathbf{a}} |M_{\mathbf{b}}(\mathbf{a}) - M'_{{\mathbf{b}}}({\mathbf{a}})||R_{\mathbf{a}}(\mathbf{b}, \mathbf{c})) \\
\leq \sum_{\mathbf{a}, \mathbf{b}, \mathbf{c}} (\sum_{\mathbf{a}} |M_{\mathbf{b}}(\mathbf{a}) - M'_{{\mathbf{b}}}({\mathbf{a}})||\sum_{\mathbf{a}} R_{\mathbf{a}}(\mathbf{b}, \mathbf{c})) \\
\leq \sum_{\mathbf{a}, \mathbf{b}, \mathbf{c}} \varepsilon_1 \sum_{\mathbf{a}} R_{\mathbf{a}}(\mathbf{b}, \mathbf{c}) \\
\leq \varepsilon_1 |\Sigma|^k
\]

\[
\sum_{\mathbf{w}} M(\mathbf{w})||R(\mathbf{w}) - R'(\mathbf{w})| \leq \sum_{\mathbf{w}} |R(\mathbf{w}) - R'(\mathbf{w})| \\
\leq \sum_{\mathbf{a}} \varepsilon_2 \\
\leq \varepsilon_2 |\Sigma|^k
\]

Finally we claim \( P'_x \) is a valid distribution satisfying \( \sum_{\mathbf{w}} P'_x(\mathbf{w}) = \sum_{\mathbf{w}} M'(\mathbf{w})R'(\mathbf{w}) = 1 \). In the bayes net factorization of \( M'(\mathbf{w})R'(\mathbf{w}) \) for every variable \( V_j \in \mathbf{V} \setminus \{X\} \) we have exactly one term in either the Bayes nets \( M' \) as \( M'(w_j | \mathbf{w}_{Pa}(j)) \) or in \( R' \) as \( R'(w_j | \mathbf{w}_{Pa}(j)) \) resp.. We first marginalize out the sink nodes of \( G \) using \( \sum_{w_j \in \Sigma} M'(w_j | \mathbf{w}_{Pa}(j)) = 1 \) or \( \sum_{w_j \in \Sigma} R'(w_j | \mathbf{w}_{Pa}(j)) = 1 \) as appropriate. Then we marginalize out the nodes which only have edges to the sink nodes and so on in the reverse topological order. When the vertex \( X \)'s turn comes we marginalize out \( \sum_{x' \in \Sigma} \) as it appears inside \( M' \).
4.2 Generation

Proof of Theorem 2.4. From Theorem 2.3, we obtain a distribution \( P'_x \) such that \( d_{TV}(P_x, P'_x) \leq \varepsilon \) and \( P'_x(w) \) for an assignment \( w = a \circ b \circ c \) (\( w_A = a, w_B = b, w_C = c \)) to \( V \setminus \{ X \} \) can be evaluated efficiently as \( P'_x(w) = M'_b(a)R'_a(b, c) \), where \( M' \) and \( R' \) are two collections of Bayes nets. \( P'_x \) as stated above is not efficiently samplable. So, we give a third distribution \( \hat{P}_x \) which can be sampled efficiently and satisfies \( d_{TV}(\hat{P}_x, P_2) \leq 7\varepsilon \). The subsequent discussion is computational where we work with the distribution \( P'_x \) and don’t need to consider \( P \).

We write \( R'_a(b, c) = R'_a(b)R'_a(c \mid b) \) where \( R'_a(c \mid b) \) is only defined for \( R'_a(b) \neq 0 \). We write \( P'_x(w) = M'_b(a)R'_a(b)\tilde{R}_a(c \mid b) \). Note that \( P'_x(a, b) = M'_b(a)R'_a(b) \) by marginalizing out \( c \) both sides. Hence we get \( P'_x(w) = P'_x(a, b)\tilde{R}_a(c \mid b) \).

Recall that \( A \) and \( B \) are small (\( \leq k \) and \( \leq kd \) resp.) while \( C \) is huge. Our approach is to learn \( P'_x(a, b) = M'_b(a)R'_a(b) \) as \( \hat{P}_x(a, b) = M'_b(a)\tilde{R}_a(b) \) and store it explicitly, while using rejection sampling for \( R'_a(c \mid b) \). Our samplable distribution is defined as \( \hat{P}_x(w) = \hat{P}_x(a, b)\tilde{R}_a(c \mid b) \). Sampling \( R'_a(c \mid b) \) would require large enough probability mass for \( R'_a(b) \) which we ensure. If \( R'_a(b) = 0 \) we also make sure \( \hat{P}_x(a, b) = 0 \). We make sure \( d_{TV}(P'_x(A, B), \hat{P}_x(A, B)) \leq 6\varepsilon \) which gives \( d_{TV}(P'_x, \hat{P}_x) \leq 6\varepsilon \) since \( R'_a(c \mid b) \) is unchanged.

Note that we already know \( M'_b(a) \) for every \( a, b \). Naively computing \( R'_a(b) \) would require an intractable marginalization. Instead, for every \( a \) we learn the distribution \( R'_a(B) \) as \( \tilde{R}_a(B) \) up to \( \varepsilon/|\Sigma|^k \) distance in \( d_{TV} \) by sampling from the known Bayes net \( R'_a(A, B) \). We use the empirical distribution of \( m_1, m_2 \) samples of \( R'_a(A, B) \) per \( a \), where \( m_1 = O((|\Sigma|^{k+2} + \log \frac{|\Sigma|^k}{\delta})/\varepsilon^2) \) and \( m_2 = O(|\Sigma|^{2kd+2k} \log \frac{|\Sigma|^{kd+k}}{\delta}/\varepsilon^2) \). \( m_1 \) samples make sure we learn \( R'_a(B) \) in \( d_{TV} \) distance \( \leq \varepsilon/|\Sigma|^k \) for all \( a \) except \( \leq \delta \) probability from a standard result. \( m_2 \) samples make sure if \( 0 < R'_a(b) < \varepsilon/|\Sigma|^{kd+k} \) for any \( a, b \), the empirical probability value \( \tilde{R}_a(b) \) is at most \( 2\varepsilon/|\Sigma|^{kd+k} \) except \( \leq \delta \) probability by Hoeffding’s bound. Subsequently we set \( \tilde{R}_a(b) = 0 \) for any \( b \) whose empirical probability is at most \( 2\varepsilon/|\Sigma|^{kd+k} \) and add the surplus probability to other \( b \)’s to make \( \tilde{R}_a(B) \) a distribution, for every \( a \). This step takes \( O(m_2|\Sigma|^k) \) time and satisfy \( d_{TV}(R'_a(B), \tilde{R}_a(B)) \leq 3\varepsilon/|\Sigma|^k \) for every \( a \). This step also makes sure if \( 0 < R'_a(b) < \varepsilon/|\Sigma|^{kd+k} \) then \( \hat{P}_x(a, b) = \tilde{R}_a(b) = 0 \) and if \( R'_a(b) = 0 \), then \( \hat{P}_x(a, b) = \tilde{R}_a(b) = 0 \) as no samples with \( B = b \) would turn up in this case.

We get \( \sum_{a, b} |P'_x(a, b) - P'_x(a, b)| = \sum_{a, b} M'_b(a)|R'_a(b) - \tilde{R}_a(b)| \leq \sum_a 3\varepsilon/|\Sigma|^k \leq 3\varepsilon. \) At this point one issue is \( \sum_{a, b} \hat{P}_x(a, b) \) might not be 1, although it lies in \([1 - 3\varepsilon, 1 + 3\varepsilon] \). So we compute \( \sum_{a, b} \hat{P}_x(a, b) \) and scale appropriately to make \( \hat{P}_x(A, B) \) a distribution which makes the \( d_{TV} \) distance at most \( \frac{3\varepsilon}{1 - 3\varepsilon} < 6\varepsilon \) for \( \varepsilon < 1/6 \).

From the triangle inequality, it follows \( d_{TV}(P_x, \hat{P}_x) \leq 7\varepsilon \). To sample from \( \hat{P}_x \) first sample \( a, b \sim \hat{P}_x(A, B) \) which we explicitly store. Subsequently we sample from the conditional distribution \( \tilde{R}_a(c \mid b) \) using rejection sampling (sampling until a \( b \) is hit). Since we make sure any item in the support of \( \hat{P}_x \) satisfy \( R'_a(b) > \varepsilon/|\Sigma|^k \), this step may need \( O(|\Sigma|^k \log \frac{1}{\varepsilon}) \) samples and \( O(n|\Sigma|^k \log \frac{1}{\varepsilon}) \) time from the Bayes net \( R'_a(A, B) \) per sample in the worst case to succeed with probability \( 1 - \delta \).

5 Lower bound

For the lower bound we use a well-known packing argument based on Fano’s inequality which says if there is a class of \( 2^K \) distributions with pairwise KL distance at most \( \beta \) then \( O(K/\beta) \) samples are
needed to identify a uniformly random distribution from the class. The KL distance between two distributions $P$ and $Q$ over a domain $[D]$ is defined as $\sum_i P(i) \ln \frac{P(i)}{Q(i)}$. The KL distance is known to satisfy certain chain rule which we use in the following proof (see eg. Lemma 6 in [Prz11]). We first recall Theorem 2.7.

**Theorem 2.7.** Fix integers $d, k \geq 1$ and a set $\Sigma$ of size $\geq 2$. For all sufficiently large $n$, there exists an ADMG $G$ with $n$ nodes and in-degree $d$ so that the following hold. $G$ contains a node $X$ such that $|\text{Pa}(X)| = d$ and $|S_1| = k$ (where $S_1$ is the $c$-component containing $X$). For any $Z \in \text{Pa}(X) \cup S_1$, there exists a causal Bayes net $P$ on $G$ over $\Sigma$-valued variables such that:

(i) For the observational distribution $P$, the marginal $P|_{\text{Pa}(X) \cup S_1 \setminus \{Z\}}$ is uniform but the marginal $P|_{\text{Pa}(X) \cup S_1}$ has mass at most $\alpha$ at some assignment.

(ii) There exists an intervention $x$ on $X$ such that learning the distribution $P_x$ up to $d_{TV}$-distance $\varepsilon$ with probability $9/10$ requires $\Omega(n|\Sigma|^{d/\alpha \varepsilon^2})$ samples from $P$.

**Proof.** We first show the lower bound where $Z$ is a parent of $X$, and $d = 2$. Later we show how to prove the full theorem.

Our ADMG on $n+2$ variables: $Z, X, Y_1, Y_2, \ldots, Y_n$ consists of $n$ triangles with $Z, X, Y_j$ for every $j$ where $Z$ is the source and $Y_j$ is the sink. Let $Y = \langle Y_1, Y_2, \ldots, Y_n \rangle$. Please refer to Figure 4a.

$Z$ is uniform over $\{0, 1\}$. $X = Z$ with probability $\alpha$ and $X = \bar{Z}$ with probability $1 - \alpha$. Thus $X, Z$ jointly satisfy $\alpha$-strong positivity. Each $Y_j \mid X, Z$ is one among the following two conditional distributions:

- $D_1 : Y_j = \text{Bern}(1/2 + \varepsilon/\sqrt{n})$ if $X \neq Z$,
  
  $Y_j = \text{Bern}(1/2)$ if $X = Z$,

- $D_2 : Y_j = \text{Bern}(1/2 - \varepsilon/\sqrt{n})$ if $X \neq Z$,
  
  $Y_j = \text{Bern}(1/2)$ if $X = Z$.

We create a class $C_\varepsilon$ of causal models using a code $C \subset \{0, 1\}^n$. This code has size $2^{\Theta(n)}$, and any two of them $c, d \in C$ satisfy the following: there are $\Theta(n)$ positions where $c$ is 1 and $d$ is 0. Showing existence of such a code is standard. Given a code as above, corresponding to every $c \in C$, we create a product distribution $Y \mid X, Z$: the 1 positions of $c$ use the distribution $D_1$ and the 0 positions of $c$ use the distribution $D_2$. Together with the distributions of $X, Z$ this defines a causal Bayes net $P_c$. 

Figure 4: ADMGs for Theorem 2.7
We first lower-bound the distance between the interventional distributions for any two members \( \mathcal{P}^c, \mathcal{P}^d \in \mathcal{C}_c \). Let \( S \) be the subset of indices from \([n]\) of size \( \Theta(n) \) where \( c \) is 1 and \( d \) is 0.

\[
d_{TV}(P_{X=1}^c(Y), P_{X=1}^d(Y)) \geq d_{TV}(P_{X=1}^c(S), P_{X=1}^d(S))
\]

With \( 1/2 \) probability, when \( Z = 1 \) every dimension of both the distributions are \( \text{Bern}(1/2) \) and therefore have \( d_{TV} = 0 \). We focus on the other case when every dimension of \( P_{X=1}^c(S) \) follows \( D_1 \) and \( P_{X=1}^d(S) \) follows \( D_2 \).

\[
d_{TV}(P_{X=1}^c(S), P_{X=1}^d(S)) = 1/2 \cdot d_{TV}(\text{Bern}(1/2 + \varepsilon/\sqrt{n})^{|S|}, \text{Bern}(1/2 - \varepsilon/\sqrt{n})^{|S|})
\]

\[
\geq 1/2 \cdot d_{TV}(\text{Bern}(1/2 + \varepsilon/\sqrt{n})^{|S|}, \text{Bern}(1/2)^{|S|})
\]

Claim 5.1. \( d_{TV}(\text{Bern}(1/2 + \varepsilon/\sqrt{n})^l, \text{Bern}(1/2)^l) \geq \Theta(\varepsilon) \) for \( l = \Theta(n) \), \( l \leq n \), \( \varepsilon \leq 1/4 \).

Proof.

\[
d_{TV}(\text{Bern}(1/2 + \varepsilon/\sqrt{n})^l, \text{Bern}(1/2)^l) = \sum_{i=0}^{l/2} \binom{l}{i} |(1/2 + \varepsilon/\sqrt{n})^i(1/2 - \varepsilon/\sqrt{n})^{l-i} - 1/2|^2
\]

\[
\geq \sum_{i=0}^{l/2} \binom{l}{i} 2^{-l}(1 - (1 + 2\varepsilon/\sqrt{n})^i(1 - 2\varepsilon/\sqrt{n})^{l-i})
\]

\[
\geq \sum_{i=0}^{l/2} \binom{l}{i} 2^{-l}(1 - \exp(2\varepsilon i/\sqrt{n})\exp(-2\varepsilon(l - i)/\sqrt{n}))
\]

\[
\geq \sum_{i=l/2 - \sqrt{l}}^{l/2} \binom{l}{i} 2^{-l}(1 - \exp(-2\varepsilon(l - 2i)/\sqrt{n}))
\]

\[
\geq \sum_{i=l/2 - \sqrt{l}}^{l/2} \binom{l}{i} 2^{-l}2\varepsilon(l/2 - i)/\sqrt{n}
\]

\[
= \sqrt{l} \sum_{j=0}^{l/2} \binom{l}{l/2 - j} 2^{-l}2\varepsilon j/\sqrt{n}
\]

The third line in the above uses the fact that in the range \( 0, 1, \ldots, l/2; 1/2^l \) is larger than the other term. The fourth line uses \( e^x \geq 1 + x \) and \( e^{-x} \geq 1 - x \). The sixth line uses \( 1 - e^{-x} \geq x/2 \) whenever \( x \leq 1 \). The ratio \( \binom{l}{i}/\binom{l}{l/2 - j} \) can be upper-bounded by \( \exp(j^2/(l/2 - j + 1)) = O(1) \) for \( 0 \leq j \leq \sqrt{l} \) and \( \binom{l}{l/2} \approx 2^l/\sqrt{l} \), which gives \( \Theta(\varepsilon) \) in the last summation.

Next we upper bound the KL distance of any two observational distributions from \( \mathcal{C}_c \). Considering the extreme case, we only upper bound the pairs \( \mathcal{P}^c \) and \( \mathcal{P}^d \) whose all the coordinates of \( Y \) are different: one is \( D_1 \), other is \( D_2 \). Note that the distributions \( P^c|_{X \cup Z} = P^d|_{X \cup Z} := P(X, Z) \) (say), which gives (by the chain rule):

16
\[
\begin{align*}
\text{KL}(P^c, P^d) &= \sum_{x,z} P(x,z) \text{KL}(P^c|Y, P^d|Y) \\
&= \sum_{x \neq z} P(x,z) \text{KL}(P^c|Y, P^d|Y) \quad \text{(as when } x = z \text{ both are } \text{Bern}(1/2)^n) \\
&= \alpha \text{KL}(P^c|Y, P^d|Y)
\end{align*}
\]

\(P^c|Y\) and \(P^d|Y\) are product distributions (in the extreme case) whose each component pairs are distributed as either \(\text{Bern}(1/2 + \varepsilon/\sqrt{n})\), \(\text{Bern}(1/2 - \varepsilon/\sqrt{n})\) or \(\text{Bern}(1/2 - \varepsilon/\sqrt{n})\), \(\text{Bern}(1/2 + \varepsilon/\sqrt{n})\).

Using additive property of KL we get \(\text{KL}(P^c, P^d) = \alpha \varepsilon^2\).

Therefore from Fano’s inequality, learning each interventional distribution up to \(\Theta(\varepsilon)\) distance with probability 2/3 requires \(\Omega(n/\alpha \varepsilon^2)\) samples.

We next improve the lower bound by \(|\Sigma|^d\) factor where \(d\) is the indegree. Please refer to Figure 4b. We pick \(\Sigma|^d\) random models from \(\mathcal{C}_\varepsilon\) and use them as the conditional distributions for \(Y \mid X, Z\). Then we create \(d\) more control variables \(W = \langle W_1, \ldots, W_d \rangle\) which are uniformly distributed over \(\Sigma|^d\) and indicates which of the hard distributions is followed by \(Y\). Any algorithm that want to learn the interventional distribution \(X = 1\) for this model in \(d_{TV}\) distance \(\varepsilon\) have to learn a constant fraction of \(|\Sigma|^d\) many hard interventional distributions from \(\mathcal{C}_\varepsilon\) in \(d_{TV}\) distance \(O(\varepsilon)\) over the choices of \(W\). We have already established that for every fixing of \(W\) learning the interventional distributions with 2/3 probability requires \(\Omega(n|\Sigma|^d/\alpha \varepsilon^2)\) samples. From Chernoff’s bound, learning \(9 \cdot |\Sigma|^d/10\) many distributions with 9/10 probability would require \(\Omega(n|\Sigma|^d/\alpha \varepsilon^2)\) samples.

We next show how to add hidden variables to the graph. Instead of \(Z\) being a parent of \(X\) in Figure 4b, suppose that \(Z\) is confounded with \(X\). That is, there is a hidden variable \(U\) that is a parent of \(X\) as well as \(Z\). Now, we can define the same causal models that we analyzed earlier, with \(U\) taking the place of the old \(Z\), and the new \(Z\) copying the value of \(U\). The analysis remains unchanged, as \(Z\) is not affected by an intervention on \(X\). Also, the degree of \(X\) and the size of the c-component can be made arbitrarily large by adding dummy variables. \(\square\)

6 Evaluation of Marginal Interventions

Here we discuss the problem of estimating \(P_x|\mathbf{F}\), i.e., the marginal interventional distribution of the intervention \(x\) to \(X\) on a subset of the observables \(\mathbf{F} \subseteq \mathbf{V}\). For ease of exposition, we can assume that the vertices of \(G\) are \(\mathbf{An}^+(\mathbf{F})\) as other variables do not play a role in \(P_x|\mathbf{F}\) and hence can be pruned out from the model; here \(\mathbf{An}^+(\mathbf{F})\) denotes the set of all observable ancestors of \(\mathbf{F}\), including \(\mathbf{F}\). Tian and Pearl [TP02a] provided an algorithm for this identification question when the ADMG satisfies Assumption 2.1 (See Theorem 4 of [TP02a]), a sufficient condition for identifiability\(^6\). Later works [SP06, HV08] generalized this result of Tian and Pearl for more general interventions, thus exhibiting a sufficient and necessary identifiability graphical condition for this problem.

We consider the following setting: Suppose \(\mathcal{P}\) is an unknown causal Bayes net over a known ADMG \(G\) on \(n\) observable variables \(\mathbf{V}\) that satisfies (Assumption 2.1) and \(\alpha\)-strong positivity with

\(^6\)Recall that [TP02a] proved Assumption 2.1 is necessary and sufficient for identifiability of \(P_x\). However to identify \(P_x|\mathbf{F}\), Assumption 2.1 was known to be only sufficient for identifiability.
For any subset \( \mathbf{F} \subseteq \mathbf{V} \) (Assumption 2.2) and let \( d \) denote the maximum in-degree of the graph \( G \), \( k \) denote the size of its largest c-component, and \( f = |\mathbf{F}| \). When the graph being referred to is unclear, we will subscript notation (eg: \( \text{Pa}_H(V) \) indicates the observable parents of \( V \) in graph \( H \)) to indicate the graph on which the operator is defined on.

We show finite sample bounds for estimating \( P_x|\mathbf{F} \) when the underlying ADMG satisfies Assumption 2.1, thus making results of [TP02a] quantitative. Estimating such causal effects under the necessary and sufficient graphical conditions of [SP06, TP02a] in the finite sample regime is an important and an interesting open question which we leave for future work. As mentioned in Section 2, the required marginal distribution \( P_x|\mathbf{F} \) can be estimated by taking \( O(|\Sigma|^f/\varepsilon^2) \) samples from the generator \( \hat{P}_x \), and we can use Theorem 2.4 to obtain the generator distribution \( \hat{P}_x \) where we require \( O(|\Sigma|^5kd^2n/\alpha\varepsilon^2) \) many samples from the observational distribution \( P \). Hence we get Corollary 2.5.

The time complexity of the algorithm (of Corollary 2.5) described above is exponential in \( f \). To handle problems that arise in practice for small \( \mathbf{F} \)’s, it is of interest to develop efficient algorithms for estimating \( P_x|\mathbf{F} \). In such cases the approach discussed above is superfluous, as the sample complexity depends linearly on \( n \), the total number of variables in the model, which could be unnecessarily large. Theorem 2.6, restated below, shows a sample and time-efficient algorithm when \( f \) is very small (e.g. constant).

**Theorem 2.6.** For any subset \( \mathbf{F} \subseteq \mathbf{V} \) with \( |\mathbf{F}| = f \), intervention \( x \) to \( X \) and parameter \( \varepsilon \in (0,1) \), there is an algorithm that takes \( m = \tilde{O}\left(\frac{|\Sigma|^{f+k(d+1)}2^{2(f+k(d+1)^2)}}{\alpha\varepsilon^5}\right) \) samples from \( P \) and runs in \( O(m[f+k(d+1)]|\Sigma|^{2(f+k(d+1)^2)}) \) time and returns an evaluator for a distribution \( \hat{P}_\mathbf{F} \) on \( \mathbf{F} \) such that \( d_{TV}(P_x|\mathbf{F}, \hat{P}_\mathbf{F}) \leq \varepsilon \).

The rest of this section is dedicated towards proving Theorem 2.6.

First let us discuss a high level idea of our algorithm for Theorem 2.6. The idea to handle cases with small \( f \) is to restrict our attention to the marginal distribution \( P(\mathbf{W}) \) over a small set of vertices \( \mathbf{W} \) and then apply Theorem 3 of [TP02a] (Theorem 3.6 here) over \( \mathbf{W} \). However restriction to \( \mathbf{W} \) could potentially modify the parent/child or c-component relationships (or both) across the vertices of \( \mathbf{W} \). Hence, to apply Theorem 3.6, the underlying causal graph over the vertices of \( \mathbf{W} \) should be obtained via a formal approach in such a way that the topological ordering and the conditional independence relations across vertices of \( \mathbf{W} \) are preserved. To do that we make use of a well-known latent projection algorithm [VP90] to reduce the given ADMG \( G \) (over observables \( \mathbf{V} \)) to a different ADMG \( H \) (over observables \( \mathbf{W} \)). A similar reduction using the latent projection [VP90] has also appeared in a slightly different context [TK18] – where the objective is to improve the efficiency of the algorithm.

We carefully choose the set \( \mathbf{W} \) and prune all the other variables \( \mathbf{V} \setminus \mathbf{W} \) from the graph \( G \), and then apply the latent projection to obtain \( H \) such that:

(A) The required causal effect is identifiable in \( H \);

(B) Theorem 2.3 can be applied by maintaining bounds on in-degree and c-component size of this new graph \( H \).

We will show that for \( \mathbf{W} = \mathbf{F} \cup \text{Pa}_G^-(\mathbf{S}_1) \) – both (A) and (B) hold. We will prove (A) while we prove (B); although (A) can easily be verified by pruning the vertices of \( \mathbf{V} \setminus \mathbf{W} \), one by one, by using Corollary 16 of [TK18]. Before we prove (B) we will first describe the reduction procedure so that the essence of the argument becomes clearer. Our reduction procedure is discussed next.
The reduction consists of two steps: The first step is to simplify the given graph $G$ to a much smaller graph $G'$ (defined over observables $W$) by ignoring all the other variables of $G$ (i.e., ignoring $V \setminus W$). By ignoring a certain variable we mean that the variable is considered to be hidden. Although the observable vertices of $G'$ is $W$, as desired, $G'$ is not an ADMG\(^7\). Since ADMGs are, in general, easy to analyze and parse through we then convert this general causal graph $G'$ to an ADMG $H$ using a known reduction technique – and this is the second step. This reduction procedure, which we call Reduction($G, W$) is formally discussed next.

6.1 Reduction: Pruning $G$ to a simpler graph $H$

Reduction($G, W$)

1. Let $G'$ be the graph obtained from $G$ by considering $V \setminus W$ as hidden variables.

2. **Projection Algorithm ($G'$ to $H$) [TP02b, VP90]**. The projection algorithm reduces the causal graph $G'$ to an ADMG $H$ by the following procedure:
   
   (a) For each observable variable $V_i \in V$ of $G'$, add an observable variable $V_i$ in $H$.
   
   (b) For each pair of observable variables $V_i, V_j \in V$, if there exists a directed edge from $V_i$ to $V_j$ in $G'$, or if there exists a directed path from $V_i$ to $V_j$ that contains only unobservable variables in $G'$, then add a directed edge from $V_i$ to $V_j$ in $H$.
   
   (c) For each pair of observable variables $V_i, V_j \in V$, if there exists an unobservable variable $U$ such that there exist two directed paths in $G'$ from $U$ to $V_i$ and from $U$ to $V_j$ such that both the paths contain only unobservable variables, then add a bidirected edge between $V_i$ and $V_j$ in $H$.

3. Return $H$

6.2 Properties of Reduction($G, W$)

It is well-known that the projection algorithm ($G'$ to $H$) [TP02b, VP90] preserves some of the important properties such as topological ordering and conditional independence relations. Before we discuss those, let us revisit the equivalent definitions of parents and c-components for general causal graphs with hidden variables.

**Definition 6.1 (Effective Parents for general causal graphs).** Given a general causal graph $G'$ and a vertex $V_j \in V$, the effective parents of $V_j$ is the set of all observable vertices $V_i$ such that either $V_i$ is a parent of $V_j$ or there exists a directed path from $V_i$ to $V_j$ that contains only unobservable variables in $G'$.

**Definition 6.2 (c-component for general causal graphs).** For a given general causal graph $G'$, two observable vertices $V_i$ and $V_j$ are related by the c-component relation if (i) there exists an unobservable variable $U$ such that $G'$ contains two paths (a) from $U$ to $V_i$; and (b) from $U$ to $V_j$, where both the paths use only unobservable variables, or (ii) there exists another vertex $V_z \in V$ such that $V_i$ and $V_z$ (and) $V_j$ and $V_z$ are related by the c-component relation.

\(^7\)Recall that an ADMG is a graph where the unobservables are root nodes and have exactly two observable children – denoted by bidirected edges.
The below lemma illustrates: “c-component that contains X remains the same in G and H.”

**Lemma 6.3.** Let \( S_1 \) denotes the c-component that contains X in G, \( W = Y \cup Pa^+_G(S_1) \) and \( H = \text{Reduction}(G, W) \). Then, (i) the c-component that contains X in H is also \( S_1 \); (ii) \( H \) satisfies Assumption 2.1; (iii) \( Pa^+_G(S_1) \) and \( Pa^+_H(S_1) \) are the same; (iv) \( H \) satisfies Assumption 2.2.

**Proof.** Let \( C \) denote the c-component of \( H \) that contains X. Note that \( S_1 \subseteq C \) since all those bidirected edges in \( G \) that forms \( S_1 \) are retained in \( H \) (because \( S_1 \subseteq W \)). Next we will prove that no other vertex of \( H \) share a bidirected edge with \( S_1 \) in \( H \). Suppose for contradiction there exists a vertex \( W_i \in W \) that share a bidirected edge with some node \( W_j \in S_1 \) in \( H \). This implies, during the reduction, there exists two paths in \( G' \) (U to \( W_i \) and U to \( W_j \)) such that all the variables included in these two paths, other than \( W_i \) and \( W_j \), are unobservables in \( G' \) which means all those vertices belong to \( V \setminus W \), a contradiction to the fact that \( Pa^+_G(S_1) \) is contained in \( W \). This proves (i). Since \( Pa^+_G(S_1) \subseteq W \) there can not exist a bidirected edge between X and a child of X in \( H \) which proves (ii).

Note that \( Pa^+_G(S_1) \subseteq Pa^+_H(S_1) \) – because \( W \) contains \( Pa^+_G(S_1) \). Now suppose, for contradiction, \( Pa^+_G(S_1) \subset Pa^+_H(S_1) \). Then during the reduction step there must have been an edge from an unobservable to \( S_1 \) in \( G' \), which can not be true as \( W \) contains \( Pa^+_G(S_1) \) and none of those variables are treated as hidden variables in the reduction. This proves (iii). Since \( S_1 \) and \( Pa^+_G(S_1) \) remains unchanged in both \( G \) and \( H \), Assumption 2.2 still holds in \( H \) which proves (iii). □

The projection algorithm (\( G' \) to \( H \)) is known to preserve the following set of properties.

- The c-components of \( H \) and \( G' \) are identical and the c-component factorization formula (Equation (20) in Lemma 2 of [TP02b]) holds even for the general causal graph (See Section 5 of [TP02b] for more details). They show this based on a known previously known reduction from \( G' \) to \( H \) [VP90]. The proof is based on the fact that for any subset \( S \subseteq V \) of observables variables, the induced subgraphs \( G'[S] \) and \( H[S] \) require the same set of conditional independence constraints.

- The effective parents (see Definition 6.1) of every observable node in \( G' \) is the same as the (observable) parent set of the corresponding node in \( H \).

- The observable vertices of \( G' \) and \( H \) are the same.

- Also, the topological ordering of the observable nodes of \( G' \) and \( H \) are the same.

As is common in the causality literature we do not use any other property of \( G' \) besides the above in our analysis, and hence it is sufficient to derive conclusions from this modified graph which contains only a small number of vertices.

### 6.2.1 Proof of Theorem 2.6

We know from Lemma 6.3 that whenever \( G \) satisfy Assumption 2.1, \( H = \text{Reduction}(G, W) \) with \( W = Y \cup Pa^+_G(S_1) \) satisfy Assumption 2.1 as well. For such graphs \( G \) and \( H \), while both satisfy Assumption 2.1, it is well-known that an equivalent statement of Theorem 3.6 directly follows from the proof of Theorem 3 of [TP02a] since their proof uses only the above mentioned properties. We will also extensively use (i) of Lemma 6.3. This results in the following theorem.
Theorem 6.4 (Theorem 3 of [TP02a] with respect to ADMG $H = \text{Reduction}(G, W)$). Let $P$ be a CBN over a causal graph $G = (V, E^{\to} \cup E^{\leftrightarrow})$, $X \in V$ be a designated variable and $Y \subseteq V \setminus X$. Let $S_1, \ldots, S_t$ be the c-components of $G$ and without loss of generality assume $X \in S_1$. Suppose that $G$ satisfies Assumption 2.1. Let $W = Y \cup \text{Pa}_G^+(S_1)$ for $Y \subseteq V \setminus \{X\}$. Let $H = \text{Reduction}(G, W)$ and let $S'_1, \ldots, S'_{t'}$ be the c-components of $H$ where without loss of generality let $S_1 = S'_1$. Then for any setting $x$ to $X$ and any assignment $t$ to $W \setminus \{X\}$ the interventional distribution $P_x(t)$ is given by:

$$P_x(t) = P_{t_{W \setminus S'_1}(t_{S'_1 \setminus \{X\}})} \cdot \prod_{j=2}^{t'} P_{t_{W \setminus (S'_1 \cup \{X\}) \circ x}}(t_{S'_j}) = \sum_{\tilde{x} \in \Sigma} Q_{S'_1}(t \circ \tilde{x}) \cdot \prod_{j=2}^{t'} Q_{S'_j}(t \circ x)$$

This proves part (A) discussed before. Next we prove part (B): where we provide bounds on the in-degree and the cardinality of c-components of $H$.

Lemma 6.5. Let $S_1$ be the c-component of $G$ that contains $X$. Let $W = Y \cup \text{Pa}_G^+(S_1)$, $H = \text{Reduction}(G, W)$ and let the c-components of $H$ are denoted by $S'_1, \ldots, S'_{t'}$ without loss of generality let $S_1 = S'_1$. Then:

1. The in-degree of $H$ is at most $f + k(d + 1)$.
2. $|S'_i| \leq f + kd$, for every $i$.

Proof. The fact that $H$ contains at most $f + k(d + 1)$ vertices provides the bound on the in-degree. We know from Lemma 6.3 that $S'_1$ is a c-component of $H$ and the remaining vertices of $H$ is $(Y \cup \text{Pa}_H^+(S'_1)) \setminus S'_1$ which is of size at most $f + kd$ (since $S_1 = S'_1$ and $\text{Pa}_G^+(S_1) = \text{Pa}_H(S_1)$).

We have now gathered the tools required to prove Theorem 2.6.

Proof of Theorem 2.6. Let $W = Y \cup \text{Pa}_G^+(S_1)$ and let $H = \text{Reduction}(G, W)$. The reduction $H = \text{Reduction}(G, W)$ can be performed using breadth first search/depth first search which can be done in time linear in the size of the input graph. We obtained an equivalent statement of Theorem 3.6 in Theorem 6.4. Also from Lemma 6.3 we know that the model over the causal graph $H$ satisfies both Assumption 2.1 and Assumption 2.2. Hence by substituting $n$ by $f + k(d + 1)$ – the cardinality of observables of $H$; $k$ by $f + kd$ – the size of the largest c-component of $H$; and $d$ by $f + k(d + 1)$ – the in-degree of $H$, into Theorem 2.3 we obtain the desired bounds on sample and time complexities. □

References

[ABDK18] Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. In Advances in Neural Information Processing Systems, pages 9447–9460, 2018.

[ASY+19] R. Agrawal, C. Squires, K.D. Yang, K. Shanmugam, and C. Uhler. Abcd-strategy: Budgeted experimental design for targeted causal structure discovery. Proceedings of Machine Learning Research, 89, Jan 2019.

[BAG16] Gilles Blondel, Marta Arias, and Ricard Gavald. Identifiability and transportability in dynamic causal networks. International Journal of Data Science and Analytics, 10 2016.
[BGMV20] Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, and N.V. Vinodchandran. Efficient distance approximation for structured high-dimensional distributions via learning. arXiv preprint, 2020. 10

[CDKS17] Clement L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing bayesian networks. In Conference on Learning Theory, pages 370–448, 2017. 10

[Das97] Sanjoy Dasgupta. The sample complexity of learning fixed-structure bayesian networks. Machine Learning, 29(2):165–180, Nov 1997. 10

[Dia16] Ilias Diakonikolas. Learning structured distributions. In Peter Bühlmann, Petros Drineas, Michael Kane, and Mark van der Laan, editors, Handbook of Big Data, chapter 15. CRC Press, Boca Raton, FL, 1 edition, 2016. 3

[GP95] David Galles and Judea Pearl. Testing identifiability of causal effects. In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, pages 185–195, 1995. 7

[Hal00] Joseph Y Halpern. Axiomatizing causal reasoning. Journal of Artificial Intelligence Research, 12:317–337, 2000. 7

[HB13] Alain Hauser and Peter Bühlmann. Jointly interventional and observational data: Estimation of interventional markov equivalence classes of directed acyclic graphs. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77, 03 2013. 7

[HEJ15] Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Do-calculus when the true graph is unknown. In Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence, pages 395–404, 2015. 7

[HR20] MA Hernan and JM Robins. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC, 2020. 2, 7

[HV07] James J Heckman and Edward J Vytlacil. Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. Handbook of econometrics, 6:4779–4874, 2007. 7

[HV08] Yimin Huang and Marco Valtorta. On the completeness of an identifiability algorithm for semi-markovian models. Annals of Mathematics and Artificial Intelligence, 54(4):363–408, 2008. 7, 17

[IR15] Guido W. Imbens and Donald B. Rubin. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, 2015. 2

[JZB19a] Amin Jaber, Jiji Zhang, and Elias Bareinboim. Causal identification under Markov equivalence: Completeness results. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 2981–2989. PMLR, 09–15 Jun 2019. 7

[JZB19b] Amin Jaber, Jiji Zhang, and Elias Bareinboim. Identification of conditional causal effects under markov equivalence. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 11512–11520. Curran Associates, Inc., 2019. 7

22
[Kel] Adam Kelleher. causality. https://github.com/akelleh/causality. Accessed: 2020-02-07.

[KJSB19] Murat Kocaoglu, Amin Jaber, Karthikeyan Shanmugam, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. In Advances in Neural Information Processing Systems, pages 14346–14356, 2019.

[KM99] Manabu Kuroki and Masami Miyakawa. Identifiability criteria for causal effects of joint interventions. Journal of the Japan Statistical Society, 29(2):105–117, 1999.

[KMR+94] Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E Schapire, and Linda Sellie. On the learnability of discrete distributions. In Proceedings of the twenty-sixth annual ACM symposium on Theory of computing, pages 273–282, 1994.

[Lew19] Arthur Lewbel. The identification zoo: Meanings of identification in econometrics. Journal of Economic Literature, 57(4):835–903, December 2019.

[Mat93] Rosa L. Matzkin. Nonparametric identification and estimation of polychotomous choice models. Journal of Econometrics, 58(1):137 – 168, 1993.

[Pea95] Judea Pearl. Causal diagrams for empirical research. Biometrika, 82(4):669–688, 1995.

[Per19] Emilija Perkovi. Identifying causal effects in maximally oriented partially directed acyclic graphs. arXiv preprint arXiv:1910.02997, October 2019.

[PR95] Judea Pearl and James Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, pages 444–453. Morgan Kaufmann Publishers Inc., 1995.

[Prz11] Pawel Przytycki, Sep 2011. From Mark Braverman’s lecture notes. https://www.cs.princeton.edu/courses/archive/fall11/cos597D/L03.pdf.

[Rob86] James Robins. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. Mathematical modelling, 7:1393–1512, 1986.

[RSR11] R. Rubin, D.S. Strayer, and E. Rubin. Rubin’s Pathology: Clinicopathologic Foundations of Medicine. Rubin’s Pathology. Wolters Kluwer Health/Lippincott Williams & Wilkins, 2011.

[SGS00] Peter Spirtes, Clark N Glymour, and Richard Scheines. Causation, prediction, and search. MIT press, 2000.

[SHW15] Amit Sharma, Jake M Hofman, and Duncan J Watts. Estimating the causal impact of recommendation systems from observational data. In Proceedings of the Sixteenth ACM Conference on Economics and Computation, pages 453–470, 2015.
[SK] Amit Sharma and Emre Kiciman. DoWhy. https://github.com/Microsoft/DoWhy/. Accessed: 2020-02-07. 7

[Sob00] Michael E. Sobel. Causal inference in the social sciences. Journal of the American Statistical Association, 95(450):647–651, 2000. 7

[SP06] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In Proceedings of the National Conference on Artificial Intelligence, volume 21, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006. 7, 17, 18

[Spi10] Peter Spirtes. Introduction to causal inference. Journal of Machine Learning Research, 11(May):1643–1662, 2010. 7

[SS16] Leonard J. Schulman and Piyush Srivastava. Stability of causal inference. In Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI’16, pages 666–675, Arlington, Virginia, United States, 2016. AUAI Press. 7

[Thw13] Peter Thwaites. Causal identifiability via chain event graphs. Artificial Intelligence, 195:291 – 315, 2013. 7

[Tia] Jin Tian. CIBN. http://web.cs.iastate.edu/~jtian/Software/CIBN.htm. Accessed: 2020-02-07. 7

[Tia02] Jin Tian. Studies in causal reasoning and learning. University of California, Los Angeles, 2002. 9

[TK17a] Santtu Tikka and Juha Karvanen. Identifying causal effects with the r package causal-effect. Journal of Statistical Software, 76, 02 2017. 7

[TK17b] Santtu Tikka and Juha Karvanen. simplifying probabilistic expressions in causal inference. Journal of Machine Learning Research, 18:1–30, 04 2017. 7

[TK18] Santtu Tikka and Juha Karvanen. Enhancing identification of causal effects by pruning. Journal of Machine Learning Research, 18:1–23, 06 2018. 7, 18

[TP02a] Jin Tian and Judea Pearl. A general identification condition for causal effects. In Proceedings of the 18th National Conference on Artificial Intelligence, pages 567–573, 2002. 3, 4, 5, 7, 10, 17, 18, 20, 21

[TP02b] Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, UAI’02, pages 519–527, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. 4, 19, 20

[TSR10] Peter Thwaites, Jim Q. Smith, and Eva Riccomagno. Causal analysis with chain event graphs. Artificial Intelligence, 174(12):889 – 909, 2010. 7

[VP90] Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence, UAI ’88, pages 69–78, Amsterdam, The Netherlands, The Netherlands, 1990. North-Holland Publishing Co. 18, 19, 20
Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pages 5541–5550, 2018.