Learning and Sampling of Atomic Interventions from Observations

Arnab Bhattacharyya [*1] **Sutanu Gayen** [*1] **Saravanan Kandasamy** [*2] **Ashwin Maran** [*3] **N. V. Vinodchandran** [*4]

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 $\mathcal{P}$ is a causal model on a set $\mathbf{V}$ 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.[1] We show that assuming that G has bounded in-degree, bounded c-components (k), 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 $d_{\mathrm{TV}}(P_x, P') \leqslant \varepsilon$ using $m = \tilde{O}(n\varepsilon^{-2})$ samples from P and $O(mn)$ time. The evaluator can return in $O(n)$ time the probability $P'(\mathbf{v})$ for any assignment $\mathbf{v}$ to $\mathbf{V}$.

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

We extend our techniques to estimate marginals $P_x|_{\mathbf{Y}}$ over a given subset $\mathbf{Y} \subseteq \mathbf{V}$ 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 ε, as well as if $k = 1$ on the strong positivity parameter.

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 (Imbens & Rubin, 2015; Pearl, 2009; Rubin et al., 2011; Hernan & Robins, 2020). In particular, (Pearl, 2009) 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

[*]Equal contribution [1]National University of Singapore [2]Cornell University [3]University of Wisconsin-Madison [4]University of Nebraska-Lincoln. Correspondence to: Arnab Bhattacharyya <arnabb@nus.edu.sg>.

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

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

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 $\mathbf{V}$ while the h additional nodes correspond to a set of h *hidden variables* $\mathbf{U}$. We assume that the observable variables take values over a finite alphabet Σ. By interpreting $\mathcal{P}$ as a standard Bayes net over $\mathbf{V} \cup \mathbf{U}$ and then marginalizing to $\mathbf{V}$, we get the observational distribution P on $\mathbf{V}$. The modularity assumption allows us to define the result of an *intervention* on $\mathcal{P}$. An intervention is specified by a subset $\mathbf{X} \subseteq \mathbf{V}$ of variables and an assignment[2] $\mathbf{x} \in \Sigma^{|\mathbf{X}|}$. In the interventional distribution, the variables $\mathbf{X}$ are fixed to $\mathbf{x}$, while each variable $W \in (\mathbf{V} \cup \mathbf{U}) \setminus \mathbf{X}$ is sampled as it would have been in the original Bayes net, according to the conditional distribution $W \mid \mathbf{Pa}(W)$, where $\mathbf{Pa}(W)$ (parents of W) consist of either variables previously sampled in the topological order of G or variables in $\mathbf{X}$ set by the intervention. The marginal of the resulting distribution to $\mathbf{V}$ is the interventional distribution denoted by $P_{\mathbf{x}}$. We sometimes also use $\mathsf{do}(\mathbf{x})$ to denote the intervention process and $P(\mathbf{V} \mid \mathsf{do}(\mathbf{x}))$ 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
$$P_x(\mathbf{v}) \stackrel{\text{def}}{=} \Pr_{\mathbf{V} \sim P_x}[\mathbf{V} = \mathbf{v}]$$
for any $\mathbf{v} \in \Sigma^n$. 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. **(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 Σ.

[2]Consistent with the convention in the causality literature, we will use a lower case letter (e.g., $\mathbf{x}$) to denote an assignment to the subset of variables corresponding to its upper case counterpart (e.g., $\mathbf{X}$).

Evaluation and generation are two very natural inference problems[3]. Indeed, the influential work of (Kearns et al., 1994) 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 (Diakonikolas, 2016) 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 & Pearl, 2002b) studied the general question of when the interventional distribution P_x is identifiable from the observational distribution P. They characterized the class $\mathcal{G}_X$ of directed acyclic graphs with hidden variables such that for any $G \in \mathcal{G}_X$, for any causal Bayes net $\mathcal{P}$ on G, and for any intervention x to X, P_x is identifiable from P. Thus, for all our work we assume that $G \in \mathcal{G}_X$, 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 & Pearl, 2002b). (Tian & Pearl, 2002b) (as well as other related work on identifiability) assumes, in addition to the underlying graph being in $\mathcal{G}_X$, 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

[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.

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 the causality literature (see (Tian & Pearl, 2002a)) 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 $\mathbf{V}$ 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 $\mathbf{V}$ gets partitioned into c-components: $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_\ell$.

Let X be a designated variable in $\mathbf{V}$. Without loss of generality, suppose $X \in \mathbf{S}_1$.

Assumption 2.1 (Identifiability with respect to X)**.** *There does not exist a path of bidirected edges between X and any child of X. Equivalently, no child of X belongs to $\mathbf{S}_1$.*

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

Assumption 2.2 (α-strong positivity with respect to X)**.** *Suppose X lies in the c-component $\mathbf{S}_1$, and let $\mathbf{Z} = \mathbf{Pa}^+(\mathbf{S}_1)$. For every assignment $\mathbf{z}$ to $\mathbf{Z}$, $P(\mathbf{Z} = \mathbf{z}) > \alpha$.*

So, if $|\mathbf{Pa}^+(\mathbf{S}_1)|$ 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 *α-strongly positive with respect to X*. More generally, if an observational distribution P satisfies $\forall \mathbf{s}, P(\mathbf{S} = \mathbf{s}) > \alpha$ for some $\alpha > 0$ and subset $\mathbf{S}$ of variables we say P is *α-strongly positive with respect to $\mathbf{S}$*.

2.1. Algorithms

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

[4]Formal definitions appear in Section 3.

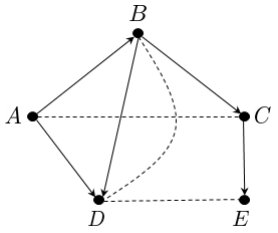


Figure 1. An acyclic directed mixed graph (ADMG) where the bidirected edges are depicted as dashed. The in-degree of the graph is 2. The c-components are $\{A, C\}$ and $\{B, D, E\}$.

We present an efficient algorithm for the evaluation and generation problems.

Theorem 2.3. *[Evaluation and Generation] For any intervention x to X and parameter $\varepsilon \in (0, 1)$, there is an algorithm that takes $m = \tilde{O}\left(\frac{|\Sigma|^{2kd} n}{\alpha^k \varepsilon^2}\right)$ samples from P, and in $O\left(mn|\Sigma|^{2kd}\right)$ time, learns a distribution $\hat{P}$ satisfying $d_{\mathrm{TV}}(P_x, \hat{P}) \leqslant \varepsilon$ such that*

- *Evaluation: Given an assignment $\mathbf{w}$ to $\mathbf{V} \setminus \{X\}$ computing $\hat{P}(\mathbf{w})$ takes $O(n|\Sigma|(kd + k))$ time*
- *Generation: Obtaining an independent sample from $\hat{P}$ takes $O(n|\Sigma|(kd + k))$ time .*

We now discuss the problem of estimating $P_x|_{\mathbf{F}}$, i.e., the marginal interventional distribution upon intervention x to X over a subset of the observables $\mathbf{F} \subseteq \mathbf{V}$. We show finite sample bounds for estimating $P_x|_{\mathbf{F}}$ when the causal Bayes net satisfies Assumption 2.1 and Assumption 2.2, thus obtaining quantitative counterparts to the results shown in (Tian & Pearl, 2002b) (See Theorem 4 of (Tian & Pearl, 2002b)). We use f to denote the cardinality of $\mathbf{F}$.

A generator for P_x obviously also gives a generator for the marginal of P_x on any subset $\mathbf{F}$. We observe that given a generator, we can also learn an approximate evaluator for the marginal of P_x on $\mathbf{F}$ sample-efficiently. This is because using $O(|\Sigma|^f/\varepsilon^2)$ samples of $\hat{P}_x$, we can learn an explicit description of $\hat{P}_x|_{\mathbf{F}}$ upto total variation distance ε with probability at least $9/10$, by simply using the empirical estimator. Since $\hat{P}_x$ is itself ε-close to P_x in total variation distance, we get an algorithm that with constant probability, returns an evaluator for a distribution that is 2ε-close to $P_x|_{\mathbf{F}}$. Summarizing:

Corollary 2.4. *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 $m = \tilde{O}\left(\frac{|\Sigma|^{2kd} n}{\alpha^k \varepsilon^2}\right)$ samples from P and in*

$O(mn|\Sigma|^{2kd})$ time returns an evaluator for a distribution $\hat{P}_x|_{\mathbf{F}}$ on $\mathbf{F}$ such that $d_{\mathrm{TV}}(P_x|_{\mathbf{F}}, \hat{P}_x|_{\mathbf{F}}) \leqslant \varepsilon$.

Note that the 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 Appendix C. Precisely, we show the following theorem:

Theorem 2.5. *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 $m = \tilde{O}\left(\frac{|\Sigma|^{5(f+k(d+1))^2}}{\alpha^k \varepsilon^2}\right)$ 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 $\tilde{P}_{\mathbf{F}}$ on $\mathbf{F}$ such that $d_{\mathrm{TV}}(P_x|_{\mathbf{F}}, \tilde{P}_{\mathbf{F}}) \leqslant \varepsilon$.*

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 α. We also explore whether Assumption 2.2 can be weakened. Since in this section, our focus is on the sample complexity instead of time complexity, we do not distinguish between evaluation and generation.

To get some intuition, consider the simple causal Bayes net depicted in Figure 2a. Here, X does not have any parents and X is not confounded with any variable. Y is a child of X, and suppose X and Y are boolean variables, where $P(X=0) = \alpha$ for some small α. Now, to estimate the interventional probability $P_{X=0}(Y=0) = P(Y=0 \mid X=0)$ to within $\pm\varepsilon$, it is well-known that $\Omega(\varepsilon^{-2})$ samples (X,Y) with $X=0$ are needed. Since $X=0$ occurs with probability α, an $\Omega(\alpha^{-1}\varepsilon^{-2})$ 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 2.6. *Fix integers $d, k \geqslant 1$ and a set Σ of size $\geqslant 2$. 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 $|\mathbf{Pa}(X)| = d$ and $|\mathbf{S}_1| = k$ (where $\mathbf{S}_1$ is the c-component containing X). For any $Z \in \mathbf{Pa}(X) \cup \mathbf{S}_1$, there exists a causal Bayes net $\mathcal{P}$ on G over Σ-valued variables such that:*

(i) For the observational distribution P, the marginal $P|_{(\mathbf{Pa}(X)\cup\mathbf{S}_1)\setminus\{Z\}}$ is uniform but the marginal $P|_{\mathbf{Pa}(X)\cup\mathbf{S}_1}$ has mass at most α at some assignment.

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

So, P must have a guarantee that its marginal on $\mathbf{Pa}(X) \cup \mathbf{S}_1$ has mass $> \alpha$ at all points in order for an algorithm to learn P_x using $O(n|\Sigma|^d/\alpha\varepsilon^2)$ samples. For comparison, our algorithm in Theorem 2.3 assume strong positivity for $\mathbf{Pa}^+(\mathbf{S}_1)$ and achieve sample complexity $O(n|\Sigma|^{2kd}/\alpha^k\varepsilon^2)$. For small values of k and d, the upper and lower bounds are close. It remains an open question to fully close the gap.

To hint towards the proof of Theorem 2.6, 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 $\mathcal{P}$ and $\mathcal{Q}$. For both P and Q, Z is uniform over $\{0,1\}$ and $X \neq Z$ with probability α. Now, suppose $P(Y=1 \mid X \neq Z) = 1/2 + \varepsilon$ and $P(Y=1 \mid X=Z) = 1/2$, while $Q(Y=1 \mid X \neq Z) = 1/2 - \varepsilon$ and $Q(Y=1 \mid X=Z) = 1/2$. Note that $P_{X=1}(Y=1) = (1+\varepsilon)/2$ while $Q_{X=1}(Y=1) = (1-\varepsilon)/2$, so that the interventional distributions are ε-far from each other. On the other hand, it can be shown using Fano's inequality that any algorithm needs to observe $\Omega(\alpha^{-1}\varepsilon^{-2})$ 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 nonparametric 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 (Pearl, 2009; Robins, 1986; Spirtes et al., 2000). 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 (Tian & Pearl, 2002b; Spirtes et al., 2000; Galles & Pearl, 1995; Halpern, 2000; Kuroki & Miyakawa, 1999; Pearl & Robins, 1995). This led to a complete algorithm[5], first by Tian and Pearl for the identifiability of atomic interventions (Tian & Pearl, 2002b) (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 (Shpitser

[5]Complete 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.

(a) Lower bound for when X is a source

(b) Lower bound for when X has a parent

Figure 2.

& Pearl, 2006) (see also (Huang & Valtorta, 2008)). 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 (Tikka & Karvanen, 2017a). This work was followed by a sequence of works (Tikka & Karvanen, 2017b; 2018) 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 (Tian; Kelleher; Sharma & Kiciman).

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 (Jaber et al., 2019a) (and (Jaber et al., 2019b)) 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 (Perkovi, 2019). Complete algorithms are also known for *dynamic causal networks*, a causal analogue for dynamic Bayesian networks that evolve over time (Blondel et al., 2016). *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 (Thwaites et al., 2010; Thwaites, 2013).

In a different line of work reported in (Schulman & Srivastava, 2016), 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 (Shpitser & Pearl, 2006). They also show that, in contrast for atomic interventions (i.e., when X is singleton) the identification algorithm of Tian and Pearl (Tian & Pearl, 2002b) is not too sensitive to changes in the input whenever Assumption 2.1 of (Tian & Pearl, 2002b) 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 (Hyttinen et al., 2015; Hauser & Bhlmann, 2013; Agrawal et al., 2019; Yang et al., 2018; Kocaoglu et al., 2019). 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 (Sharma et al., 2015), computational sciences (Spirtes, 2010), social and behavioral sciences (Sobel, 2000), econometrics (Heckman & Vytlacil, 2007; Matzkin, 1993; Lewbel, 2019), and epidemiology (Hernan & Robins, 2020).

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 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 (Acharya et al., 2018) 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. (Jung et al., 2020) gave a weighting-based estimator for efficiently estimating an identifiable causal effect using finite samples.

3. Preliminaries

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

The variables in this paper take values in a finite set Σ. 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 vari-

ation distance is denoted by $d_{\mathrm{TV}}(P,Q)$ and is given by $d_{\mathrm{TV}}(P,Q) := \frac{1}{2}\sum_{i\in[D]}|P(i)-Q(i)|$. The KL distance between them is defined as $\sum_i P(i)\ln\frac{P(i)}{Q(i)}$. Pinsker's inequality says $d_{\mathrm{TV}}(P,Q) \leqslant \sqrt{2\mathrm{KL}(P,Q)}$.

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* $\langle \mathbf{V}, G, \{\mathbf{Pr}[V_i \mid \mathbf{pa}(V_i)] : V_i \in \mathbf{V}, \mathbf{pa}(V_i) \in \Sigma^{|\mathbf{Pa}(V_i)|}\}\rangle$ *where: (i)* $\mathbf{V} = (V_1,\ldots,V_n)$ *is a set of variables over alphabet* Σ*, (ii)* G *is a directed acyclic graph with* n *nodes corresponding to the elements of* $\mathbf{V}$*, and (iii)* $\mathbf{Pr}[V_i \mid \mathbf{pa}(V_i)]$ *is the conditional distribution of variable* V_i *given that its parents* $\mathbf{Pa}(V_i)$ *in* G *take the values* $\mathbf{pa}(V_i)$*.*

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

$$P(\mathbf{v}) = \prod_{V_i\in\mathbf{V}} \mathbf{Pr}[v_i \mid \mathbf{Pa}(V_i) = \mathbf{v}_{\mathbf{Pa}(V_i)}].$$

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

Causality. We describe Pearl's (1995) notion of causality. Central to his formalism is the notion of an *intervention*. Given an observable variable set $\mathbf{V}$ and a subset $\mathbf{X} \subset \mathbf{V}$, an intervention $\mathsf{do}(\mathbf{x})$ is the process of fixing the set of variables $\mathbf{X}$ to the values $\mathbf{x}$. The *interventional distribution* $P_{\mathbf{x}}$ is the distribution on $\mathbf{V}$ after setting $\mathbf{X}$ to $\mathbf{x}$. 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* $\langle \mathbf{V}, \mathbf{U}, G, \{\mathbf{Pr}[V_i \mid \boldsymbol{\pi}(V_i)] : V_i \in \mathbf{V}, \boldsymbol{\pi}(V_i) \in \Sigma^{|\mathbf{\Pi}(V_i)|}\}, \{\mathbf{Pr}[\mathbf{U}]\}\rangle$*, where (i)* $\mathbf{V} = (V_1,\ldots,V_n)$ *and* $\mathbf{U}$ *are the tuples of observable and hidden variables respectively, (ii)* G *is a directed acyclic graph on* $\mathbf{V}\cup\mathbf{U}$*, (iii)* $\mathbf{Pr}[V_i \mid \boldsymbol{\pi}(V_i)]$ *is the conditional probability distributions of* $V_i \in \mathbf{V}$ *given that its parents* $\mathbf{\Pi}(V_i) \in \mathbf{V}\cup\mathbf{U}$ *take the values* $\boldsymbol{\pi}(V_i)$*, and (iv)* $\mathbf{Pr}[\mathbf{U}]$ *is the distribution of the hidden variables* $\mathbf{U}$*.* G *is said to be the* causal graph *corresponding to* $\mathcal{P}$*.*

Such a causal Bayes net $\mathcal{P}$ *defines a unique interventional distribution* $P_{\mathbf{x}}$ *for every subset* $\mathbf{X} \subseteq \mathbf{V}$ *(including* $\mathbf{X} = \emptyset$*) and assignment* $\mathbf{x} \in \Sigma^{|\mathbf{X}|}$*, as follows. For all* $\mathbf{v} \in \Sigma^{|\mathbf{V}|}$*:*

$$P_{\mathbf{x}}(\mathbf{v}) = \begin{cases} \sum_{\mathbf{u}} \prod_{V_i\in\mathbf{V}\setminus\mathbf{X}} \mathbf{Pr}[v_i \mid \mathbf{\Pi}(V_i) = \mathbf{v}_{\mathbf{\Pi}(V_i)}]\cdot\mathbf{Pr}[\mathbf{u}] \\ \text{if } \mathbf{v} \text{ is consistent with } \mathbf{x} \\ 0 \quad \text{otherwise.} \end{cases}$$

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

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 $E^{\leftrightarrow}$. Given such an ADMG G, for any $\mathbf{S} \subseteq \mathbf{V}$, $\overline{\mathbf{S}}$ denotes the complement set $\mathbf{V}\setminus\mathbf{S}$, $\mathbf{Pa}(\mathbf{S})$ denotes the parents of $\mathbf{S}$ according to the directed edges of G, i.e., $\mathbf{Pa}(\mathbf{S}) = \cup_{X\in\mathbf{S}}\{Y \in \mathbf{V} : (Y,X) \in E^{\rightarrow}\}$. We also define: $\mathbf{Pa}^{+}(\mathbf{S}) = \mathbf{Pa}(\mathbf{S})\cup\mathbf{S}$ and $\mathbf{Pa}^{-}(\mathbf{S}) = \mathbf{Pa}(\mathbf{S})\setminus\mathbf{S}$. The bidirected edges are used to define c-components:

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

Since a c-component forms an equivalence relation, the set of all c-components forms a partition of $\mathbf{V}$, the observable vertices of G. Let $\mathbf{S}_1 \cup \mathbf{S}_2 \cup \cdots \cup \mathbf{S}_\ell$ denote the partition of $\mathbf{V}$ into the c-components of G.

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

$$Q_{\mathbf{S}}(\mathbf{v}) = P_{\mathbf{v}_{\overline{\mathbf{S}}}}(\mathbf{v}_{\mathbf{S}}).$$

Clearly, for every $\mathbf{v}_{\overline{\mathbf{S}}}$*,* $Q_{\mathbf{S}}$ *is a distribution over* $\Sigma^{|\mathbf{S}|}$*.*

For $\mathbf{Y} \subseteq \mathbf{V}$, the induced subgraph $G[\mathbf{Y}]$ is the subgraph obtained by removing the vertices $\mathbf{V}\setminus\mathbf{Y}$ and their corresponding edges from G.

The following lemma is used heavily in this work.

Lemma 3.5 (Corollary 1 of (Tian, 2002))**.** *Let* $\mathcal{P}$ *be a causal Bayes net on* $G = (\mathbf{V}, E^{\rightarrow}\cup E^{\leftrightarrow})$*. Let* $\mathbf{S}_1,\ldots,\mathbf{S}_\ell$ *be the c-components of* G*. Then for any* $\mathbf{v}$ *we have:*

(i) $P(\mathbf{v}) = \prod_{i=1}^{\ell} Q_{\mathbf{S}_i}(\mathbf{v})$*.*

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

$$Q_{\mathbf{S}_j}(\mathbf{v}) = \prod_{i:V_i\in\mathbf{S}_j} P(v_i \mid v_1,\ldots,v_{i-1}).$$

(iii) Furthermore, each factor $P(v_i \mid v_1,\ldots,v_{i-1})$ *can be expressed as:*

$$P(v_i \mid v_1,\ldots,v_{i-1}) = P(v_i \mid \mathbf{v}_{\mathbf{Pa}^{+}(\mathbf{T}_i)\cap[i-1]})$$

where $\mathbf{T}_i$ *is the c-component of* $G[V_1,\ldots,V_i]$ *that contains* V_i*.*

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

Theorem 3.6 (Theorem 3 of (Tian & Pearl, 2002b))**.** *Let $\mathcal{P}$ be a causal Bayes net over $G = (\mathbf{V}, E^{\rightarrow} \cup E^{\leftrightarrow})$ and $X \in \mathbf{V}$ be a variable. Let $\mathbf{S}_1, \ldots, \mathbf{S}_\ell$ be the c-components of G and assume $X \in \mathbf{S}_1$ 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 $\mathbf{w}$ to $\mathbf{V} \setminus \{X\}$, the interventional distribution $P_x(\mathbf{w})$ is given by:*

$$P_x(\mathbf{w}) = P_{\mathbf{w}_{\mathbf{V}\setminus\mathbf{S}_1}}(\mathbf{w}_{\mathbf{S}_1\setminus\{X\}}) \cdot \prod_{j=2}^{\ell} P_{\mathbf{w}_{\mathbf{V}\setminus(\mathbf{S}_j\cup\{X\})}\circ x}(\mathbf{w}_{\mathbf{S}_j})$$
$$= \sum_{x'\in\Sigma} Q_{\mathbf{S}_1}(\mathbf{w}\circ x') \cdot \prod_{j=2}^{\ell} Q_{\mathbf{S}_j}(\mathbf{w}\circ x)$$

4. Efficient Estimation

Let $\mathcal{P}$ be a causal Bayes net over a causal graph $G = (\mathbf{V}, E^{\rightarrow} \cup E^{\leftrightarrow})$. 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. Before we proceed to our algorithms for interventional distributions, we will first present an algorithm for *learning the observational distribution* $P(\mathbf{V})$. Our approach is to then view the causal Bayes net as a regular Bayes net over observable variables and use the learning algorithm for Bayes nets. From Lemma 3.5, we can write the observational distribution $P(\mathbf{V})$ as:

$$P(\mathbf{V}) = \prod_{i=1}^{n} P(V_i \mid \mathbf{Z}_i) \tag{1}$$

where $\mathbf{Z}_i \subseteq \{V_1, \ldots, V_{i-1}\}$ is the set of 'effective parents' of V_i of size at most $kd + k$. Here k is the maximum c-component size and d is the maximum in-degree. Therefore the observational distribution 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 (Canonne et al., 2017), starting from Dasgupta's (1997) early work. In this work, we will use the following learning result described in (Bhattacharyya et al., 2020).

Theorem 4.1 ((Bhattacharyya et al., 2020))**.** *There is an algorithm that on input parameters $\varepsilon, \delta \in (0,1)$ and samples from an unknown Bayes net P over Σ^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 $d_{\mathrm{TV}}(P, \hat{P}) \leqslant \varepsilon$ with probability $\geqslant 1 - \delta$.*

From the above discussion we get the following corollary.

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

In the next subsection we design our evaluation and generation algorithms.

4.1. Evaluation and Sampling of P_x

In this section we will prove Theorem 2.3 for learning P_x. Let the index of $X \in \mathbf{V}$ be t in the topological ordering i.e. $X = V_t$. Let $\mathbf{S}_1$ be the c-component containing X. According to Equation (1), the observational distribution $P(\mathbf{V})$ factorizes as a Bayes net into factors of the form $P(V_i \mid \mathbf{Z}_i)$ where $\mathbf{Z}_i$ are the effective dependants of V_i.

We note that the interventional distribution P_x can be represented as a marginal distribution of a different Bayes net $D_x(\mathbf{V})$ based on the following observation that uses Theorem 3.6. To obtain this representation of P_x, consider the Bayes net factorization of $P(\mathbf{V})$. Replace all the factors $P(V_i \mid \mathbf{Z}_i)$ satisfying $V_i \notin \mathbf{S}_1$ and $X \in \mathbf{Z}_i$ by $P(V_i \mid \mathbf{Z}_i \setminus \{X\}, X = x)$. In other words, each of these factors now does not use the variable X and instead uses the constant $X = x$ (which is the intervention) as parent. All the other factors, including $P(V_t \mid \mathbf{Z}_t)$, remain the same as in $P(\mathbf{V})$. The marginal distribution of D_x on $\mathbf{V} \setminus \{X\}$ is exactly P_x. This is illustrated in Figure 3a and Figure 3b.

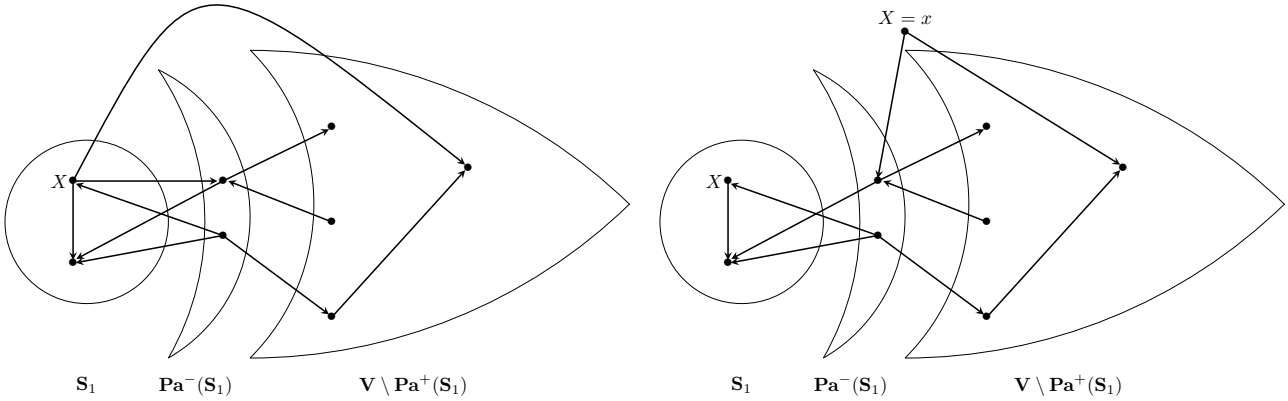
More formally, let $\mathbf{W} := \mathbf{V} \setminus \{X\}$ and $\mathbf{w}$ be an arbitrary assignment to it and let $X = V_t$. Using Theorem 3.6, P_x can be factorized as follows:

$$P_x(\mathbf{w}) = \left(\sum_{x'\in\Sigma} \left(\prod_{V_i\in\mathbf{S}_1} P((\mathbf{w}\circ x')_{V_i} \mid (\mathbf{w}\circ x')_{\mathbf{Z}_i})\right)\right) \prod_{V_i\notin\mathbf{S}_1} P(\mathbf{w}_{V_i} \mid (\mathbf{w}\circ x)_{\mathbf{Z}_i}) \tag{2}$$

where $\mathbf{Z}_i$ is the effective parents of V_i from Equation (1). So, we start with the factorization of Equation (1) for the assignment $\mathbf{w} \circ x$, then replace all occurrences of x with x' in $\mathbf{Z}_i \cup \{V_i\}$ of $P(V_i \mid \mathbf{Z}_i)$ for every $V_i \in \mathbf{S}_1$ and taking a summation over all possible values of $x' \in \Sigma$.

In a view to learn P_x, we learn the Bayes net distribution:

$$D_x(\mathbf{V}) = \prod_{\substack{V_i\in\mathbf{S}_1 \vee \\ X\notin\mathbf{Z}_i}} P(V_i \mid \mathbf{Z}_i) \prod_{\substack{V_i\notin\mathbf{S}_1 \wedge \\ X\in\mathbf{Z}_i}} P(V_i \mid \mathbf{Z}_i \setminus \{X\}, x). \tag{3}$$



(a) The Bayes net $P(\mathbf{V})$.

(b) The Bayes net $D_x(\mathbf{V})$. Variables in $\mathbf{V}\setminus\mathbf{S}_1$ which originally depended on the variable X in P now instead depend on the constant $X = x$.

Figure 3.

So, we start with the factorization of Equation (1) and replace all occurrences of the variable X with the constant x which appear in the factors outside of $\mathbf{S}_1$. D_x is a well-defined distribution as $\sum_{\mathbf{v}} D_x(\mathbf{v}) = 1$ by marginalizing out variables one after another in the reverse topological order, starting from the sink nodes. Learning D_x suffices since its marginal on $\mathbf{V} \setminus \{X\}$ is exactly $P_x(\mathbf{W})$:

$$P_x(\mathbf{W}) = \sum_{x' \in \Sigma} D_x(\mathbf{W} \circ x').$$

We rewrite Equation (3) as:

$$D_x(\mathbf{V}) = \prod_{V_i} D_x(V_i \mid \mathbf{Z}'_i) \tag{4}$$

where $\mathbf{Z}'_i = \mathbf{Z}_i$ and $D_x(V_i \mid \mathbf{Z}'_i) = P(V_i \mid \mathbf{Z}'_i)$ for $V_i \in \mathbf{S}_1 \vee X \notin \mathbf{Z}_i$; and $\mathbf{Z}'_i = \mathbf{Z}_i \setminus \{X\}$ and $D_x(V_i \mid \mathbf{Z}'_i) = P(V_i \mid \mathbf{Z}'_i, X = x)$ otherwise.

We use a KL local subadditivity result for Bayes nets from (Canonne et al., 2017). For a Bayes net R, a vertex i and an assignment $\mathbf{a}$ to its parents, let $\Pi[i, \mathbf{a}]$ denote the event that the parents of a variable i is $\mathbf{a}$ and let $R(i \mid \mathbf{a})$ denote the conditional distribution of variable i when its parents are $\mathbf{a}$.

Theorem 4.3 ((Canonne et al., 2017))**.** *Let R, S be two Bayes nets over a common graph. Then*

$$\mathrm{KL}(R, S) \leqslant \sum_i \sum_{\mathbf{a}} R(\Pi[i, \mathbf{a}])\ \mathrm{KL}(R(i \mid \mathbf{a}), S(i \mid \mathbf{a}))$$

We also need the following result for learning the local distributions in KL distance.

Theorem 4.4 ((Kamath et al., 2015))**.** *Let D be an unknown distribution over Σ. Suppose we take z samples from D and define the add-1 empirical distribution $D'(i) = (z_i + 1)/(z + |\Sigma|)$ where z_i is the number of occurrences of item $i \in \Sigma$. Then $\mathbf{E}[\mathrm{KL}(D, D')] \leqslant (|\Sigma| - 1)/(z + 1)$.*

We are trying to learn D_x but we have only sample access to P. The following lemma relates the p.m.f.s of D_x and P which we use later.

Lemma 4.5. *Let $\mathbf{w}$ be an assignment to $\mathbf{V}\setminus\{X\}$ and let x' and x be two assignments to X. Suppose P be α-strongly positive w.r.t. $\mathbf{Pa}^+(\mathbf{S}_1)$. Then the following holds:*

1. $P(\mathbf{w} \circ x) \geqslant \alpha^k D_x(\mathbf{w} \circ x')$
2. $P(\mathbf{w} \circ x) \geqslant \frac{\alpha^k}{|\Sigma|} D_x(\mathbf{w})$
3. $P(\mathbf{w}) \geqslant \frac{\alpha^k}{|\Sigma|} D_x(\mathbf{w})$

Proof. Let $\mathbf{v} = \mathbf{w} \circ x$ and $\mathbf{v}' = \mathbf{w} \circ x'$.

$$\begin{aligned}
\frac{P(\mathbf{w} \circ x)}{D_x(\mathbf{w} \circ x')} &= \frac{\prod_i P(v_i \mid \mathbf{v}_{\mathbf{Z}_i})}{\prod_{V_i \in \mathbf{S}_1} P(v'_i \mid \mathbf{v}'_{\mathbf{Z}_i}) \prod_{V_i \notin \mathbf{S}_1} P(v_i \mid \mathbf{v}_{\mathbf{Z}_i})} \\
&= \frac{\prod_{V_i \in \mathbf{S}_1} P(v_i \mid \mathbf{v}_{\mathbf{Z}_i})}{\prod_{V_i \in \mathbf{S}_1} P(v'_i \mid \mathbf{v}'_{\mathbf{Z}_i})} \\
&\geqslant \prod_{V_i \in \mathbf{S}_1} P(v_i \mid \mathbf{v}_{\mathbf{Z}_i}) \\
&\geqslant \prod_{V_i \in \mathbf{S}_1} P(v_i, \mathbf{v}_{\mathbf{Z}_i}) \\
&\geqslant \alpha^k
\end{aligned}$$

The first line uses Equation (1) and Equation (3). The fourth line follows from $P(A \mid B) \geqslant P(A \mid B)P(B) =$

Algorithm 1 Learning D_x

Input: Samples from P, parameters m, t
Output: A Bayes net $\hat{D}_x$ according to the factorization of Equation (4)
Get m samples from P
for every vertex $V_i \in \mathbf{S}_1$ **do**
  for every fixing $\mathbf{Z}_i = \mathbf{a}$, where $\mathbf{Z}_i$ are the effective parents of V_i **do**
    $\hat{D}_x(V_i \mid \mathbf{Z}'_i = \mathbf{a}) \leftarrow$ the add-1 empirical distribution (see Theorem 4.4) at node i in the subset of samples where $\mathbf{Z}_i = \mathbf{a}$
  end for
end for
for every vertex $V_i \in \mathbf{V} \setminus \mathbf{S}_1$ **do**
  for every fixing $\mathbf{Z}_i \setminus \{X\} = \mathbf{a}$, where $\mathbf{Z}_i$ are the effective parents of V_i **do**
    if $X \in \mathbf{Z}_i$ **then**
      $N_{i,\mathbf{a}} \leftarrow$ the number of samples with $\mathbf{Z}_i \setminus \{X\} = \mathbf{a}$ and $X = x$
      if $N_{i,\mathbf{a}} \geqslant t$ **then**
        $\hat{D}_x(V_i \mid \mathbf{Z}_i \setminus \{X\} = \mathbf{a}) \leftarrow$ the add-1 empirical distribution at node i in the subset of samples where $\mathbf{Z}_i \setminus X = \mathbf{a}$ and $X = x$
      else
        $\hat{D}_x(V_i \mid \mathbf{Z}_i \setminus \{X\} = \mathbf{a}) \leftarrow$ the uniform distribution over Σ
      end if
    else
      $N_{i,\mathbf{a}} \leftarrow$ the number of samples with $\mathbf{Z}_i = \mathbf{a}$
      if $N_{i,\mathbf{a}} \geqslant t$ **then**
        $\hat{D}_x(V_i \mid \mathbf{Z}_i = \mathbf{a}) \leftarrow$ the add-1 empirical distribution at node i in the subset of samples where $\mathbf{Z}_i = \mathbf{a}$
      else
        $\hat{D}_x(V_i \mid \mathbf{Z}_i = \mathbf{a}) \leftarrow$ the uniform distribution over Σ
      end if
    end if
  end for
end for

$P(AB)$ for any two events A and B. The last line follows since for $V_i \in \mathbf{S}_1$, $\{V_i\} \cup \mathbf{Z}_i \subseteq \mathbf{Pa}^+(\mathbf{S}_1)$ and P is α-strongly positive w.r.t. the later.

Part 2 follows by marginalization of Part 1 over all possible $x' \in \Sigma$. Part 3 trivially follows from Part 2. □

Finally we give Algorithm 1 along with Lemma 4.6 for learning D_x as a Bayes net according to the factorization of Equation (4). Its proof can be found in Appendix A.

Lemma 4.6. *Let $D_x(\mathbf{V})$ be the Bayes net as defined in Equation (4). Then Algorithm 1 takes $\tilde{O}\left(\frac{n|\Sigma|^{2kd}}{\alpha^k \varepsilon^2}\right)$ samples and $\tilde{O}\left(\frac{n^2|\Sigma|^{4kd}}{\alpha^k \varepsilon^2}\right)$ time and returns a Bayes net $\hat{D}_x(\mathbf{V})$ such that $d_{\mathrm{TV}}(D_x, \hat{D}_x) \leqslant \varepsilon$ with probability at least 3/4.*

We repeat Algorithm 1 independently $O(\log \frac{1}{\delta})$ times to achieve $(1 - \delta)$ success probability. This gives us Theorem 2.3. See Appendix A for the details.

Acknowledgements The works of AB and SG are supported in part by AB's Start-up Grant WBS R252000A33133. The work of SK is supported by a Cornell University Grant. The work of VV is supported by NSF Awards 1849048 and 1934884.

References

Acharya, J., Bhattacharyya, A., Daskalakis, C., and Kandasamy, S. Learning and testing causal models with interventions. In *Advances in Neural Information Processing Systems*, pp. 9447–9460, 2018. 5

Agrawal, R., Squires, C., Yang, K., Shanmugam, K., and Uhler, C. Abcd-strategy: Budgeted experimental design for targeted causal structure discovery. *Proceedings of Machine Learning Research*, 89, Jan 2019. 5

Bhattacharyya, A., Gayen, S., Meel, K. S., and Vinodchandran, N. Efficient distance approximation for structured high-dimensional distributions via learning. *arXiv preprint*, 2020. 7, 13

Blondel, G., Arias, M., and Gavald, R. Identifiability and transportability in dynamic causal networks. *International Journal of Data Science and Analytics*, 10 2016. doi: 10.1007/s41060-016-0028-8. 5

Canonne, C. L., Diakonikolas, I., Kane, D. M., and Stewart, A. Testing bayesian networks. In *Conference on Learning Theory*, pp. 370–448, 2017. 7, 8

Dasgupta, S. The sample complexity of learning fixed-structure bayesian networks. *Machine Learning*, 29(2): 165–180, Nov 1997. ISSN 1573-0565. doi: 10.1023/A:1007417612269. 7

Diakonikolas, I. Learning structured distributions. In Bühlmann, P., Drineas, P., Kane, M., and van der Laan, M. (eds.), *Handbook of Big Data*, chapter 15. CRC Press, Boca Raton, FL, 1 edition, 2016. 2

Galles, D. and Pearl, J. Testing identifiability of causal effects. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pp. 185–195, 1995. ISBN 1-55860-385-9. 4

Halpern, J. Y. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12:317–337, 2000. 4

Hauser, A. and Bhlmann, P. 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. doi: 10.1111/rssb.12071. 5

Heckman, J. J. and Vytlacil, E. J. Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of econometrics*, 6:4779–4874, 2007. 5

Hernan, M. and Robins, J. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020. 1, 5

Huang, Y. and Valtorta, M. On the completeness of an identifiability algorithm for semi-markovian models. *Annals of Mathematics and Artificial Intelligence*, 54(4):363–408, 2008. 5, 16

Hyttinen, A., Eberhardt, F., and Järvisalo, M. Do-calculus when the true graph is unknown. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pp. 395–404, 2015. 5

Imbens, G. W. and Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751. 1

Jaber, A., Zhang, J., and Bareinboim, E. 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*, pp. 2981–2989. PMLR, 09–15 Jun 2019a. 5

Jaber, A., Zhang, J., and Bareinboim, E. Identification of conditional causal effects under markov equivalence. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 11512–11520. Curran Associates, Inc., 2019b. 5

Jung, Y., Tian, J., and Bareinboim, E. Estimating causal effects using weighting-based estimators. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 10186–10193. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6579. 5

Kamath, S., Orlitsky, A., Pichapati, D., and Suresh, A. T. On learning distributions from their samples. In Grnwald, P., Hazan, E., and Kale, S. (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1066–1100, Paris, France, 03–06 Jul 2015. PMLR. URL http://proceedings.mlr.press/v40/Kamath15.html. 8

Kearns, M., Mansour, Y., Ron, D., Rubinfeld, R., Schapire, R. E., and Sellie, L. On the learnability of discrete distributions. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pp. 273–282, 1994. 2

Kelleher, A. causality. https://github.com/akelleh/causality. Accessed: 2020-02-07. 5

Kocaoglu, M., Jaber, A., Shanmugam, K., and Bareinboim, E. Characterization and learning of causal graphs with latent variables from soft interventions. In *Advances in Neural Information Processing Systems*, pp. 14346–14356, 2019. 5

Kuroki, M. and Miyakawa, M. Identifiability criteria for causal effects of joint interventions. *Journal of the Japan Statistical Society*, 29(2):105–117, 1999. 4

Lewbel, A. The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903, December 2019. doi: 10.1257/jel.20181361. 5

Matzkin, R. L. Nonparametric identification and estimation of polychotomous choice models. *Journal of Econometrics*, 58(1):137 – 168, 1993. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(93)90116-M. 5

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. 6

Pearl, J. *Causality*. Cambridge university press, 2009. 1, 4

Pearl, J. and Robins, J. Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 444–453. Morgan Kaufmann Publishers Inc., 1995. 4

Perkovi, E. Identifying causal effects in maximally oriented partially directed acyclic graphs. *arXiv preprint arXiv:1910.02997*, October 2019. 5

Przytycki, P., Sep 2011. From Mark Braverman's lecture notes. https://www.cs.princeton.edu/courses/archive/fall11/cos597D/L03.pdf. 14

Robins, J. 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. 4

Rubin, R., Strayer, D., and Rubin, E. *Rubin's Pathology: Clinicopathologic Foundations of Medicine*. Rubin's Pathology. Wolters Kluwer Health/Lippincott Williams & Wilkins, 2011. ISBN 9781605479682. 1

Schulman, L. J. and Srivastava, P. Stability of causal inference. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pp. 666–675, Arlington, Virginia, United States, 2016. AUAI Press. ISBN 978-0-9966431-1-5. 5

Sharma, A. and Kiciman, E. DoWhy. https://github.com/Microsoft/DoWhy/. Accessed: 2020-02-07. 5

Sharma, A., Hofman, J. M., and Watts, D. J. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pp. 453–470, 2015. 5

Shpitser, I. and Pearl, J. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, pp. 1219–1226, 2006. 4, 5, 16

Sobel, M. E. Causal inference in the social sciences. *Journal of the American Statistical Association*, 95(450):647–651, 2000. ISSN 01621459. 5

Spirtes, P. Introduction to causal inference. *Journal of Machine Learning Research*, 11(May):1643–1662, 2010. 5

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, prediction, and search*. MIT press, 2000. 4

Thwaites, P. Causal identifiability via chain event graphs. *Artificial Intelligence*, 195:291 – 315, 2013. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.09.003. 5

Thwaites, P., Smith, J. Q., and Riccomagno, E. Causal analysis with chain event graphs. *Artificial Intelligence*, 174(12):889 – 909, 2010. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2010.05.004. 5

Tian, J. CIBN. http://web.cs.iastate.edu/~jtian/Software/CIBN.htm. Accessed: 2020-02-07. 5

Tian, J. *Studies in causal reasoning and learning*. University of California, Los Angeles, 2002. 6

Tian, J. and Pearl, J. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pp. 519–527, San Francisco, CA, USA, 2002a. Morgan Kaufmann Publishers Inc. ISBN 1-55860-897-4. 3, 17, 18

Tian, J. and Pearl, J. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pp. 567–573, 2002b. 2, 3, 4, 5, 7, 16, 17, 18

Tikka, S. and Karvanen, J. Identifying causal effects with the r package causaleffect. *Journal of Statistical Software*, 76, 02 2017a. doi: 10.18637/jss.v076.i12. 5

Tikka, S. and Karvanen, J. simplifying probabilistic expressions in causal inference. *Journal of Machine Learning Research*, 18:1–30, 04 2017b. 5

Tikka, S. and Karvanen, J. Enhancing identification of causal effects by pruning. *Journal of Machine Learning Research*, 18:1–23, 06 2018. 5, 17

Verma, T. and Pearl, J. Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '88, pp. 69–78, Amsterdam, The Netherlands, The Netherlands, 1990. North-Holland Publishing Co. ISBN 0-444-88650-8. 17, 18

Yang, K., Katcoff, A., and Uhler, C. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pp. 5541–5550, 2018. 5