Learning and Testing Causal Models with Interventions

Jayadev Acharya*
School of ECE
Cornell University
acharya@cornell.edu

Arnab Bhattacharyya†
CSA Department
Indian Institute of Science
arnabb@iisc.ac.in

Constantinos Daskalakis‡
EECS
MIT
costis@csail.mit.edu

Saravanan Kandasamy§
CSA Department
Indian Institute of Science
saravan.tuty@gmail.com

May 25, 2018

Abstract

We consider testing and learning problems on causal Bayesian networks as defined by Pearl [Pea09]. Given a causal Bayesian network $\mathcal{M}$ on a graph with n discrete variables and bounded in-degree and bounded "confounded components", we show that $O(\log n)$ interventions on an unknown causal Bayesian network $\mathcal{X}$ on the same graph, and $\tilde{O}(n/\varepsilon^2)$ samples per intervention, suffice to efficiently distinguish whether $\mathcal{X} = \mathcal{M}$ or whether there exists some intervention under which $\mathcal{X}$ and $\mathcal{M}$ are farther than ε in total variation distance. We also obtain sample/time/intervention efficient algorithms for: (i) testing the identity of two unknown causal Bayesian networks on the same graph; and (ii) learning a causal Bayesian network on a given graph. Although our algorithms are non-adaptive, we show that adaptivity does not help in general: $\Omega(\log n)$ interventions are necessary for testing the identity of two unknown causal Bayesian networks on the same graph, even adaptively. Our algorithms are enabled by a new subadditivity inequality for the squared Hellinger distance between two causal Bayesian networks.

1 Introduction

A central task in statistical inference is learning properties of a high-dimensional distribution over some variables of interest given observational data. However, probability distributions only capture the association between variables of interest and may not suffice to predict what the consequences would be of setting some of the variables to particular values. A standard example illustrating the point is this: From observational data, we may learn that atmospheric air pressure and the readout of a barometer are correlated. But can we predict whether the atmospheric pressure would stay the same or go up if the barometer readout was forcefully increased by moving its needle?

Such issues are at the heart of *causal inference*, where the goal is to learn a *causal model* over some variables of interest, which can predict the result of *external interventions* on the variables. For example, a

*Supported by a Cornell University startup grant.
†Partially supported by DST Ramanujan grant DSTO1358 and DRDO Frontiers Project DRDO0687.
‡Partially supported by NSF CCF-1650733, CCF-1617730, CCF-1741137.
§Partially supported by DRDO Frontiers Project DRDO0687.

causal model on two variables of interest X and Y need not only determine conditional probabilities of the form $\mathbf{Pr}[Y \mid X = x]$, but also *interventional probabilities* $\mathbf{Pr}[Y \mid do(X = x)]$ where, following Pearl's notation [Pea09], $do(X = x)$ means that X has been forced to take the value x by an external action. In our previous example, $\mathbf{Pr}[\text{Pressure} \mid do(\text{Barometer} = b)] = \mathbf{Pr}[\text{Pressure}]$ but $\mathbf{Pr}[\text{Barometer} \mid do(\text{Pressure} = p)] \neq \mathbf{Pr}[\text{Barometer}]$, reflecting that the atmospheric pressure causes the barometer readout, not the other way around.

Causality has been the focus of extensive study, with a wide range of analytical frameworks proposed to capture causal relationships and perform causal inference. A prevalent class of causal models are *graphical causal models*, going back to Wright [Wri21] who introduced such models for path analysis, and Haavelmo [Haa43] who used them to define structural equation models. Today, graphical causal models are widely used to represent causal relationships in a variety of ways [SDLC93, GC99, Pea09, SGS00, Nea04, KF09].

In our work, we focus on the central model of *causal Bayesian networks* (CBNs) [Pea09, SGS00, Nea04]. Recall that a (standard) Bayesian network is a distribution over several random variables that is associated with a directed acyclic graph. The vertices of the graph are the random variables over which the distribution is defined, and the graph describes conditional independence properties of the distribution. In particular, every variable is independent of its non-descendants, conditioned on the values of its parents in the graph. A CBN is also associated with a directed acyclic graph (DAG) whose vertices are the random variables on which the distribution is defined. However, a CBN is not a single distribution over these variables but the collection of all possible interventional distributions, defined by setting any subset of the variables to any set of values. In particular, every vertex is both a variable V and a *mechanism* to generate the value of V given the values of the parent vertices, and the interventional distributions are defined in terms of these mechanisms.

We allow CBNs to contain both observable and unobservable (hidden) random variables. Importantly, we allow *unobservable confounding variables*. These are variables that are not observable, yet they are ancestors of at least two observable variables. These are especially tricky in statistical inference, as they may lead to spurious associations.

1.1 Our Contributions

Consider the following situations:

1. An engineer designs a large circuit using a circuit simulation program and then builds it in hardware. The simulator predicts relationships between the voltages and currents at different nodes of the circuit. Now, the engineer would like to verify whether the simulator's predictions hold for the real circuit by doing a limited number of experiments (e.g., holding some voltages at set levels, cutting some wires, etc.). If not, then she would want to learn a model for the system that has sufficiently good accuracy.

2. A biologist is studying the role of a set of genes in migraine. He would like to know whether the mechanisms relating the products of these genes are approximately the same for patients with and without migraine. He has access to tools (e.g., CRISPR-based gene editing technologies [DPL+16]) that generate data for gene activation and knockout experiments.

Motivated by such scenarios, we study the problems of hypothesis testing and learning CBNs when both observational and interventional data are available. The main highlight of our work is that we prove bounds on the number of samples, interventions, and time steps required by our algorithms.

To define our problems precisely, we need to specify what we consider to be a good approximation of a causal model. Given $\varepsilon \in (0, 1)$, we say that two causal models $\mathcal{M}$ and $\mathcal{N}$ on a set of variables

$\mathbf{V} \cup \mathbf{U}$ (observable and unobservable resp.) are *ε-close* (denoted $\Delta(\mathcal{M}, \mathcal{N}) \leqslant \varepsilon$) if for every subset $\mathbf{S}$ of $\mathbf{V}$ and assignment $\mathbf{s}$ to $\mathbf{S}$, performing the same intervention $do(\mathbf{S} = \mathbf{s})$ to both $\mathcal{M}$ and $\mathcal{N}$ leads to the two interventional distributions being ε-close to each other in total variation distance. Otherwise, the two models are said to be *ε-far* and $\Delta(\mathcal{M}, \mathcal{N}) > \varepsilon$.

Thus, two models $\mathcal{M}$ and $\mathcal{N}$ are close according to the above definition if there is *no* intervention which can make the resulting distributions differ significantly. This definition is motivated by the philosophy articulated by Pearl (pp. 414, [Pea09]) that "causation is a summary of behavior under intervention". Intuitively, if there is some intervention that makes $\mathcal{M}$ and $\mathcal{N}$ behave differently, then $\mathcal{M}$ and $\mathcal{N}$ do not describe the same causal process. Without having any prior information about the set of relevant interventions, we adopt a worst-case view and simply require that causal models $\mathcal{M}$ and $\mathcal{N}$ behave similarly for every intervention to be declared close to each other.[1]

The *goodness-of-fit testing* problem can now be described as follows. Suppose that a collection $\mathbf{V} \cup \mathbf{U}$ (observable and unobservable resp.) of n random variables are causally related to each other. Let $\mathcal{M}$ be a hypothesized causal model for $\mathbf{V} \cup \mathbf{U}$ that we are given explicitly. Suppose that the true model to describe the causal relationships is an unknown $\mathcal{X}$. Then, the goodness-of-fit testing problem is to distinguish between: (i) $\mathcal{X} = \mathcal{M}$, versus (ii) $\Delta(\mathcal{X}, \mathcal{M}) > \varepsilon$, by sampling from and experimenting on $\mathbf{V}$, i.e. forcing some variables in $\mathbf{V}$ to certain values and sampling from the thus intervened upon distribution.

We study goodness-of-fit testing assuming $\mathcal{X}$ and $\mathcal{M}$ are causal Bayesian networks over a known DAG G. Given a DAG G, CBN $\mathcal{M}$ and $\varepsilon > 0$, we denote the corresponding goodness-of-fit testing problem $\mathsf{CGFT}(G, \mathcal{M}, \varepsilon)$. For example, the engineer above, who wants to determine whether the circuit behaves as the simulation software predicts, is interested in the problem $\mathsf{CGFT}(G, \mathcal{M}, \varepsilon)$ where $\mathcal{M}$ is the simulator's prediction, G is determined by the circuit layout, and ε is a user-specified accuracy parameter. Here is our theorem for goodness-of-fit testing.

Theorem 1.1 (Goodness-of-fit Testing – Informal)**.** *Let G be a DAG on n vertices with bounded in-degree and bounded "confounded components." Let $\mathcal{M}$ be a given CBN over G. Then, there exists an algorithm solving $\mathsf{CGFT}(G, \mathcal{M}, \varepsilon)$ that makes $O(\log n)$ interventions, takes $\tilde{O}(n/\varepsilon^2)$ samples per intervention and runs in time $\tilde{O}(n^2/\varepsilon^2)$. Namely, the algorithm gets access to a CBN $\mathcal{X}$ over G, accepts with probability $\geqslant 2/3$ if $\mathcal{X} = \mathcal{M}$ and rejects with probability $\geqslant 2/3$ if $\Delta(\mathcal{X}, \mathcal{M}) > \varepsilon$.*

By "confounded component" in the above statement, we mean a *c-component* in G, as defined in Definition 2.10. Roughly, a c-component is a maximal set of observable vertices that are pairwise connected by paths of the form $V \leftarrow U \rightarrow V \leftarrow U \rightarrow V \leftarrow \cdots \rightarrow V$ where V and U correspond to observable and unobservable variables respectively. The decomposition of CBNs into c-components has been important in earlier work [TP02] and continues to be an important structural property here.

We can use our techniques to extend Theorem 1.1 in several ways:

(1) In the *two-sample testing* problem for causal models, the tester gets access to two unknown causal models $\mathcal{X}$ and $\mathcal{Y}$ on the same set of variables $\mathbf{V} \cup \mathbf{U}$ (observable and unobservable resp.). For a given $\varepsilon > 0$, the goal is to distinguish between (i) $\mathcal{X} = \mathcal{Y}$ and (ii) $\Delta(\mathcal{X}, \mathcal{Y}) > \varepsilon$ by sampling from and intervening on $\mathbf{V}$ in both $\mathcal{X}$ and $\mathcal{Y}$.
We solve the two-sample testing problem when the inputs are two CBNs over the same DAG G in n variables; for a given $\varepsilon > 0$ and DAG G, call the problem $\mathsf{C2ST}(G, \varepsilon)$. Specifically, we show an algorithm to solve $\mathsf{C2ST}(G, \varepsilon)$ that makes $O(\log n)$ interventions on the input models $\mathcal{X}$ and $\mathcal{Y}$, uses

[1] To quote Pearl again, "It is the nature of any causal explanation that its utility be proven not over standard situations but rather over novel settings that require innovative manipulations of the standards." (pp. 219, [Pea09]).

$\tilde{O}(n/\varepsilon^2)$ samples per intervention and runs in time $\tilde{O}(n^2/\varepsilon^2)$, when G has bounded in-degree and c-component size.[2]

(2) For the $\mathsf{C2ST}(G, \varepsilon)$ problem, the requirement that G be fully known is rather strict. Instead, suppose the common graph G is unknown and only bounds on its in-degree and maximum c-component size are given. For example, the biologist above who wants to test whether certain causal mechanisms are identical for patients with and without migraine can reasonably assume that the underlying causal graph is the same (even though he doesn't know what it is exactly) and that only the strengths of the relationships may differ between subjects with and without migraine. For this problem, we obtain an efficient algorithm with nearly the same number of samples and interventions as above.

(3) The problem of *learning* a causal model can be posed as follows: the learning algorithm gets access to an unknown causal model $\mathcal{X}$ over a set of variables $\mathbf{V} \cup \mathbf{U}$ (observable and unobservable resp.), and its objective is to output a causal model $\mathcal{N}$ such that $\Delta(\mathcal{X}, \mathcal{N}) \leqslant \varepsilon$.
We consider the problem $\mathsf{CL}(G, \varepsilon)$ of learning a CBN over a known DAG G on the observable and unobservable variables. For example, this is the problem facing the engineer above who wants to learn a good model for his circuit by conducting some experiments; the DAG G in this case is known from the circuit layout. Given a DAG G with bounded in-degree and c-component size and a parameter $\varepsilon > 0$, we design an algorithm that on getting access to a CBN $\mathcal{X}$ defined over G, makes $O(\log n)$ interventions, uses $\tilde{O}(n^2/\varepsilon^4)$ samples per intervention, runs in time $\tilde{O}(n^3/\varepsilon^4)$, and returns an oracle $\mathcal{N}$ that can efficiently compute $P_{\mathcal{X}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{T} = \mathbf{t})]$ for any $\mathbf{T} \subseteq \mathbf{V}$ and $\mathbf{t} \in \Sigma^{|\mathbf{T}|}$ with error at most ε in TV distance.

The sample complexity of our testing algorithms matches the state-of-the-art for testing identity of (standard) Bayes nets [DP17, CDKS17]. Designing a goodness-of-fit tester using $o(n)$ samples is a very interesting challenge and seems to require fundamentally new techniques.

We also show that the number of interventions for $\mathsf{C2ST}(G, \varepsilon)$ and $\mathsf{CL}(G, \varepsilon)$ is nearly optimal, even in its dependence on the in-degree and c-component size, and even when the algorithms are allowed to be adaptive. By 'adaptive' we mean the algorithms are allowed to choose the future interventions based on the samples observed from the past interventions. Specifically,

Theorem 1.2. *There exists a causal graph G on n vertices, with maximum in-degree at most d and largest c-component size at most ℓ, such that $\Omega(|\Sigma|^{\ell d - 2} \log n)$ interventions are necessary for any algorithm (even adaptive) that solves $\mathsf{C2ST}(G, \varepsilon)$ or $\mathsf{CL}(G, \varepsilon)$.*

1.2 Related Work

1.2.1 Causality

As mentioned before, there is a huge and old literature on causality, for both testing causal relationships and inferring causal graphs that is impossible to detail here. Below, we point out some representative directions of research that are relevant to our work. This discussion is far from exhaustive, and the reader is encouraged to pursue the references cited in the mentioned works.

Most work on statistical tests for causal models has been in the parametric setting. *Structural equation models* have traditionally been tested for goodness-of-fit by comparing observed and predicted covariance matrices [BL92]. Another class of tests that has been proposed assumes that the causal factors and the noise factors are conditionally independent. In the *additive noise model* [HJM+09, PJS11, ZPJS12, SSS+17], each

[2] Of course, it is allowed for the two networks to be different subgraphs of G. So, $\mathcal{X}$ could be defined by the graph G_1 and $\mathcal{Y}$ by G_2. Our result holds when $G_1 \cup G_2$ is a DAG with bounded in-degree and c-component size.

variable is the sum of a (non-linear) function of its parent variables and independent noise, often assumed to be Gaussian. This point of view has been refined into an information-geometric criterion in [JMZ+12]. In the non-parametric setting, which is the concern of this paper, Tian and Pearl [TP02] show how to derive functional constraints from causal Bayesian graphs that give equality and inequality constraints among the (distributions of) observed variables, not just conditional independence relations. Kang and Tian [KT06] derive such functional constraints on interventional distributions. Although these results yield non-trivial constraints, they are valid for any model that respects a particular graph and it is not clear how to use them for testing goodness-of-fit with statistical guarantees.

Learning in the context of causal inference has been extensively studied. To the best of our knowledge, though, most previous work is on learning only the causal graph, whereas our objective is to learn the entire causal model (i.e., the set of all interventional distributions). Pearl and Verma [PV95, VP92] investigated the problem of finding a causal graph with hidden variables that is consistent with a given list of conditional independence relations in observational data. In fact, there may be a large number of causal graphs that are consistent with a given set of conditional independence relations. [SGS00, ARSZ05], and Zhang [Zha08] (building on the FCI algorithm [SMR99]) has given a complete and sound algorithm for recovering a representative of the equivalence class consistent with a set of conditional independence relations.

Subsequent work considered the setting when both observational and interventional data are available. This setting has been a recent focus of study [HB12a, WSYU17, YKU18], motivated by advances in genomics that allow high-resolution observational and interventional data for gene expression using flow cytometry and CRISPR technologies [SPP+05, MBS+15, DPL+16]. When there are no confounding variables, Hauser and Bühlmann [HB12b], following up on work by Eberhardt and others [EGS05, Ebe07], find the information-theoretically minimum number of interventions that are sufficient to identify[3] the underlying causal graph and provide a polynomial time algorithm to find such a set of interventions. A recent paper [KDV17] extends the work of [HB12b] to minimize the total cost of interventions where each vertex is assigned a cost. Another work by Shanmugam et al. [SKDV15] investigates the problem of learning causal graphs without confounding variables using interventions on sets of small size. In the presence of confounding variables, there are several works which aim to learn the causal graph from interventional data (e.g., [MMLM06, HEH13]). In particular, a recent work of Kacaoglu et al. [KSB17] gives an efficient randomized algorithm to learn a causal graph with confounding variables while minimizing the number of interventions from which conditional independence relations are obtained.

All the works mentioned above assume access to an oracle that gives conditional independence relations between variables in the observed and interventional distributions. This is clearly a problematic assumption because it implicitly requires unbounded training data. For example, Scheines and Spirtes [SS08] have pointed out that measurement error, quantization and aggregation can easily alter conditional independence relations. The problem of developing finite sample bounds for testing and learning causal models has been repeatedly posed in the literature. The excellent survey by Guyon, Janzing and Schölkopf [GJS10] on causality from a machine learning perspective underlines the issue as one of the "ten open problems" in the area. To the best of our knowledge, our work is the first to show finite sample complexity and running time bounds for inference problems on causal Bayesian networks.

An application of our learning algorithm is to the problem of *transportability*, studied in [BP13, SP08, LH13, PB11, BP12], which refers to the notion of transferring causal knowledge from a set of source domains to a target domain to identify causal effects in the target domain, when there are certain commonalities between the source and target domains. Most work in this area assume the existence of an algorithm that

[3] More precisely, the goal is to discover the causal graph given the conditional independence relations satisfied by the interventional distributions.

learns the set of *all* interventions, that is the complete specification of the the source domain model. Our learning algorithm can be used for this purpose; it is efficient in terms of time, interventions, and sample complexity, and it learns each intervention distribution to error at most ε.

1.2.2 Distribution Testing and Learning

There is a vast literature on testing and learning high dimensional distributions in the statistics, and information theory literature, and more recently in computer science with a focus on the computational efficiency of solving such problems. We will not be able to cover and do justice to all of these works in this section. However, we will provide pointers to some of the resources, and also discuss some of the recent progress that is the most closely related to the work we present here.

In the distribution learning and testing framework, the closest to our work is learning and testing graphical models. The seminal work of Chow-Liu [CL68] considered the problem of learning tree-structured graphical models. Motivated by applications across many fields, the problem of learning graphical models from samples has gathered recent interest. Of particular interest is the apparent gap between the sample complexity and computational complexity of learning graphical models. [AKN06, BMS08] provided algorithms for learning bounded degree graphical models with polynomial sample and time complexity. A lower bound on the sample complexity that grows exponentially with the degree, and only logarithmically with the number of dimensions was provided by [SW12], and recent works [Bre15, VMLC16, KM17] have proposed algorithms with near optimal sample complexity, and polynomial running time for learning Ising models.

Sample and computational complexity of testing graphical models has been studied recently, in [CDKS17] for testing Bayesian Networks, and in [DDK18] for testing Ising models. Given sample access to an unknown Bayesian Network, or Ising model, they study the sample complexity, and computation complexity of deciding whether the unknown model is equal to a known fixed model (hypothesis testing).

The problem of testing and learning distribution properties has itself received wide attention in statistics with a history of over a century [Fis25, LR06, CT06]. In these fields, the emphasis is on asymptotic analysis characterizing the convergence rates, and error exponents, as the number of samples tends to infinity. A recent line of work originating from [GR00, BFR$^+$00] focuses on *sublinear* algorithms where the goal is to design algorithms with the number of samples that is smaller than the domain size (e.g., [Can15, Gol17], and references therein).

While most of these results are for learning and testing low dimensional (usually one dimensional) distributions, there are some notable exceptions. Testing for properties such as independence, and monotonicity in high dimensions have been considered recently [BFRV11, ADK15, DK16]. These results show that the optimal sample complexity for testing these properties grows exponentially with the number of dimensions. A line of recent work [DP17, CDKS17, DDK17, DDK18] overcomes this barrier by utilizing additional structure in the high-dimensional distribution induced by Bayesian network or Markov Random Field assumptions.

1.3 Overview of our Techniques

In this section, we give an overview of the proof of Theorem 1.1 and the lower bound construction. We start by making a well-known observation [TP02, VP90] that CBNs can be assumed to be over a particular class of DAGs known as *semi-Markovian causal graphs*. A semi-Markovian causal graph is a DAG where every vertex corresponding to an unobservable variable is a root and has exactly two children, both observable. More details of the correspondence are given in Appendix B.

In a semi-Markovian causal graph, two observable vertices V_1 and V_2 are said to be connected by a bi-directed edge if there is a common unobservable parent of V_1 and V_2. Each connected component of the graph restricted to bi-directed edges is called a *c-component*. The decomposition into c-components gives very useful structural information about the causal model. In particular, a fact that is key to our whole analysis is that if $\mathcal{N}$ is a semi-Markovian Bayesian network on observable and unobservable variables $\mathbf{V} \cup \mathbf{U}$ with c-components $\mathbf{C}_1, \ldots, \mathbf{C}_p$, then for any $\mathbf{v} \in \Sigma^{|\mathbf{V}|}$:

$$P_{\mathcal{N}}[\mathbf{v}] = \prod_{i=1}^{p} P_{\mathcal{N}}[\mathbf{c}_i \mid do(\mathbf{V} \setminus \mathbf{C}_i = \mathbf{v} \setminus \mathbf{c}_i)] \tag{1}$$

where Σ is the alphabet set, $\mathbf{c}_i$ is the restriction of $\mathbf{v}$ to $\mathbf{C}_i$ and $\mathbf{v} \setminus \mathbf{c}_i$ is the restriction of $\mathbf{v}$ to $\mathbf{V} \setminus \mathbf{C}_i$. Moreover, one can write a similar formula (Lemma 2.12) for an interventional distribution on $\mathcal{N}$ instead of the observable distribution $P_{\mathcal{N}}[\mathbf{v}]$.

The most direct approach to test whether two causal Bayes networks $\mathcal{X}$ and $\mathcal{Y}$ are identical is to test whether each interventional distribution is identical in the two models. This strategy would require $(|\Sigma|+1)^n$ many interventions, each on a variable set of size $O(n)$, where n is the total number of observable vertices. To reduce the number of interventions as well as the sample complexity, a natural approach, given (1) and its extension to interventional distributions, is to test for identity between each pair of "local" distributions

$$P_{\mathcal{X}}[\mathbf{S} \mid do(\mathbf{v} \setminus \mathbf{s})] \qquad \text{and} \qquad P_{\mathcal{Y}}[\mathbf{S} \mid do(\mathbf{v} \setminus \mathbf{s})]$$

for every subset $\mathbf{S}$ of a c-component $\mathbf{C}$ and assignment $\mathbf{v} \setminus \mathbf{s}$ to $\mathbf{V} \setminus \mathbf{S}$. We assume that each c-component is bounded, so each local distribution has bounded support. Moreover, using the conditional independence properties of Bayesian networks, note that in each local distribution, we only need to intervene on observable parents of $\mathbf{S}$ that are outside $\mathbf{S}$, not on all of $\mathbf{V} \setminus \mathbf{S}$.

Through a probabilistic argument, we efficiently find a *small set* $\mathbf{I}$ of *covering interventions*, which are defined as a set of interventions with the following property: For every subset $\mathbf{S}$ of a c-component and for every assignment $\mathbf{pa}(\mathbf{S})$ to the observable parents of $\mathbf{S}$, there is an intervention $I \in \mathbf{I}$ that does not intervene on $\mathbf{S}$ and sets the parents of $\mathbf{S}$ to exactly $\mathbf{pa}(\mathbf{S})$. Our test performs all the interventions in $\mathbf{I}$ on both $\mathcal{X}$ and $\mathcal{Y}$ and hence can observe each of the local distributions $P_{\mathcal{X}}[\mathbf{S} \mid do(\mathbf{pa}(\mathbf{S}))]$ and $P_{\mathcal{Y}}[\mathbf{S} \mid do(\mathbf{pa}(\mathbf{S}))]$. What remains is to bound $\Delta(\mathcal{X}, \mathcal{Y})$ in terms of the distances between each pair of local distributions.

To that end, we develop a subadditivity theorem about CBNs, and this is the main technical contribution of our upper bound results. We show that if each pair of local distributions is within distance γ in *squared Hellinger* distance, then for any intervention I, applying I to $\mathcal{X}$ and $\mathcal{Y}$ results in distributions that are within $O(n\gamma)$ distance in squared Hellinger distance, assuming bounded in-degree and c-component size of the underlying graph. A bound on the total variation distance between the interventional distributions and hence $\Delta(\mathcal{X}, \mathcal{Y})$ follows. The subadditivity theorem is inspired from [DP17], where they showed that for Bayes networks, "closeness of local marginals implies closeness of the joint distribution". Our result is in a very different set-up, where we prove "closeness of local interventions implies closeness of any joint interventional distribution", and requires a new proof technique. We relax the squared Hellinger distance between the interventional distributions as the objective of a minimization program in which the constraints are that each pair of local distributions is γ-close in squared Hellinger distance. By a sequence of transformations of the program, we lower bound its objective in terms of γ, thus proving our result. In the absence of unobservable variables, the analysis becomes much simpler and is sketched in Appendix A.

Regarding the lower bound, we prove that the number of interventions required by our algorithms are indeed necessary for any algorithm that solves $\mathsf{C2ST}(G, \varepsilon)$ or $\mathsf{CL}(G, \varepsilon)$, even if the algorithms are provided

with infinite samples/time. For any algorithm that fails to perform some local intervention I, we provide a construction of two models which do not agree on I and agree on all other interventions. Our construction is designed in such a way that it allows adaptive algorithms. The idea is to show an adversary that, for each intervention, reveals a distribution to the algorithm. Towards the end, when the algorithm fails to perform some local intervention I, we can show a construction of two models such that: i) both the models do not agree on I, and the total variation distance between the interventional distributions is equal to one; ii) and for all other interventions, the interventional distributions revealed by the adversary match with the corresponding distributions on both the models. This, together with a probabilitic argument, shows the existence of a causal graph that requires sufficiently large number of interventions to solve $\mathsf{C2ST}(G, \varepsilon)$ and $\mathsf{CL}(G, \varepsilon)$.

1.4 Future Directions

We hope that this work paves the way for future research on designing efficient algorithms with bounded sample complexity for learning and testing causal models. For the sake of concreteness, we list a few open problems.

- Interventional experiments are often expensive or infeasible, so one would like to deduce causal models from observational data alone. In general, this is impossible. However, in *identifiable* causal Bayesian networks (see [Tia02]), one can identify causal effects from observational data alone. **Is there an efficient algorithm to learn an identifiable interventional distribution from samples?**[4]
- A deficiency of our work is that we assume the underlying causal graph is fully known. **Can our learning algorithm be extended to the setting where the hypothesis only consists of some limited information about the causal graph (e.g., in-degree, c-component size) instead of the whole graph?** In fact, it is open how to efficiently learn the distribution given by a Bayesian network based on samples from it, if we don't have access to the underlying graph [DP17, CDKS17].
- Our goodness-of-fit algorithm might reject even when the input $\mathcal{X}$ is very close to the hypothesis $\mathcal{M}$. **Is there a *tolerant* goodness-of-fit tester that accepts when $\Delta(\mathcal{X}, \mathcal{M}) \leqslant \varepsilon_1$ and rejects when $\Delta(\mathcal{X}, \mathcal{M}) > \varepsilon_2$ for $0 < \varepsilon_1 < \varepsilon_2 < 1$?** Our current analysis does not extend to a tolerant tester. The same question holds for two-sample testing.
- In many applications, causal models are described in terms of *structural equation models*, in which each variable is a deterministic function of its parents as well as some stochastic error terms. **Design sample and time efficient algorithms for testing and learning structural equation models.** Other questions such as evaluating *counterfactual* queries or doing *policy analysis* (see Chapter 7 of [Pea09]) also present interesting algorithmic problems.

2 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

[4] Schulman and Srivastava [SS16] have shown that under adversarial noise, there exist causal Bayesian networks on n nodes where estimating an identifiable intervention to precision d requires precision $d + \exp(n^{0.49})$ in the estimates of the probabilities of observed events. However, this instability is likely due to the adversarial noise and does not preclude an efficient sampling-based algorithm, especially if we assume a balancedness condition as in [CDKS17].

variables (sets of variables), e.g., a is the value of A and $\mathbf{b}$ is the value of the set of variables $\mathbf{B}$. The variables in this paper take values in a discrete set Σ. We use $[n]$ to denote $\{1, 2, \ldots, n\}$.

Probability and Statistics. The total variation (TV) distance between distributions P and Q over the same set $[D]$ is $\delta_{TV}(P, Q) := \frac{1}{2}\sum_{i\in[D]} |P(i) - Q(i)|$. The squared Hellinger distance (given in (9)) and the total variation distance are related by the following.

Lemma 2.1 (Hellinger vs total variation)**.** *The Hellinger distance and the total variation distance between two distributions P and Q are related by the following inequality:*

$$H^2(P, Q) \leqslant \delta_{TV}(P, Q) \leqslant \sqrt{2H^2(P, Q)}.$$

The problem of two-sample testing for discrete distributions in Hellinger distance, and learning with respect to total variation distance has been studied in the literature, and the following two lemmas state two results we use. Let P and Q denote distributions over a domain of size D.

Lemma 2.2 (Hellinger Test, [DK16])**.** *Given $\tilde{O}(\min(D^{2/3}/\varepsilon^{8/3}, D^{3/4}/\varepsilon^2))$ samples from each unknown distributions P and Q, we can distinguish between $P = Q$ vs $H^2(P, Q) \geqslant \varepsilon^2$ with probability at least $2/3$. This probability can be boosted to $1 - \delta$ at a cost of an additional $O(\log(1/\delta))$ factor in the sample complexity. The running time of the algorithm is quasi-linear in the sample size.*

Lemma 2.3 (Learning in TV distance, folklore (e.g. [DL12]))**.** *For all $\delta \in (0, 1)$, the empirical distribution $\hat{P}$ computed using $\Theta\left(\frac{D}{\varepsilon^2} + \frac{\log\frac{1}{\delta}}{\varepsilon^2}\right)$ samples from P satisfies $H^2(P, \hat{P}) \leqslant \delta_{TV}(P, \hat{P}) \leqslant \varepsilon$, with probability at least $1 - \delta$.*

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

Definition 2.4. *A* Bayesian Network *(*BN*) $\mathcal{N}$ 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) $\mathbf{V}$ is a set of variables over alphabet Σ, (ii) G is a directed acyclic graph with 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 $\mathcal{N} = \langle \mathbf{V}, G, \{\mathbf{Pr}[V_i \mid \mathbf{pa}(V_i)]\}\rangle$ defines a unique probability distribution $P_{\mathcal{N}}$ over $\Sigma^{|\mathbf{V}|}$, as follows. For all $\mathbf{v} \in \Sigma^{|\mathbf{V}|}$,

$$P_{\mathcal{N}}[\mathbf{v}] = \prod_{V_i \in \mathbf{V}} \mathbf{Pr}[v_i \mid \mathbf{pa}(V_i)].$$

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

Conditional independence relations in graphical models are captured by the following definitions.

Definition 2.5. *Given a DAG G, a (not necessarily directed) path p in G is said to be* blocked *by a set of nodes $\mathbf{Z}$, if (i) p contains a chain node B ($A \to B \to C$) or a fork node B ($A \leftarrow B \to C$) such that $B \in \mathbf{Z}$ (or) (ii) p contains a collider node B ($A \to B \leftarrow C$) such that $B \notin \mathbf{Z}$ and no descendant of B is in $\mathbf{Z}$.*

Definition 2.6 (d-separation)**.** *For a given DAG G on $\mathbf{V}$, two disjoint sets of vertices $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ are said to be* d-separated *by $\mathbf{Z}$ in G, if every (not necessarily directed) path in G between $\mathbf{X}$ and $\mathbf{Y}$ is blocked by $\mathbf{Z}$.*

Lemma 2.7 (Graphical criterion for independence)**.** *For a given* BN *$\mathcal{N} = \langle \mathbf{V}, G, \{\mathbf{Pr}[V_i \mid \mathbf{pa}(V_i)]\}\rangle$ and $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{V}$, if $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$ in G, then $\mathbf{X}$ is* independent *of $\mathbf{Y}$ given $\mathbf{Z}$ in $P_{\mathcal{N}}$, denoted by $[\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}]$ in $P_{\mathcal{N}}$.*

2.1 Causality

We describe Pearl's notion of causality from [Pea95]. Central to his formalism is the notion of an *intervention*. Given a variable set $\mathbf{V}$ and a subset $\mathbf{X} \subset \mathbf{V}$, an intervention $do(\mathbf{x})$ is the process of fixing the set of variables $\mathbf{X}$ to the values $\mathbf{x}$. The *interventional distribution* $\mathbf{Pr}[\mathbf{V} \mid do(\mathbf{x})]$ is the distribution on $\mathbf{V}$ after setting $\mathbf{X}$ to $\mathbf{x}$. As discussed in the introduction, an intervention is quite different from conditioning.

Another important component of Pearl's formalism is that some variables may be unobservable. The unobservable variables can neither be observed nor be intervened. We partition our variable set into two sets $\mathbf{V}$ and $\mathbf{U}$, where the variables in $\mathbf{V}$ are *observable* and the variables in $\mathbf{U}$ are *unobservable*. Given a directed acyclic graph H on $\mathbf{V} \cup \mathbf{U}$ and a subset $\mathbf{X} \subseteq (\mathbf{V} \cup \mathbf{U})$, we use $\mathbf{\Pi}_H(\mathbf{X})$, $\mathbf{Pa}_H(\mathbf{X})$, $\mathbf{An}_H(\mathbf{X})$, and $\mathbf{De}_H(\mathbf{X})$ to denote the set of all parents, observable parents, observable ancestors and observable descendants respectively of $\mathbf{X}$, excluding $\mathbf{X}$, in H. When the graph H is clear, we may omit the subscript. As usual, small letters, $\boldsymbol{\pi}(\mathbf{X})$, $\mathbf{pa}(\mathbf{X})$, $\mathbf{an}(\mathbf{X})$ and $\mathbf{de}(\mathbf{X})$ are used to denote their corresponding values. And, we use $H_{\overline{\mathbf{X}}}$ and $H_{\underline{\mathbf{X}}}$ to denote the graph obtained from H by removing the incoming edges to $\mathbf{X}$ and outgoing edges from $\mathbf{X}$ respectively.

Definition 2.8 (Causal Bayesian Network)**.** *A* causal Bayesian network *(*CBN*) is a collection of interventional distributions that can be defined in terms of a tuple* $\langle \mathbf{V}, \mathbf{U}, G, \{\mathbf{Pr}[V_i \mid \boldsymbol{\pi}(V_i)] : V_i \in \mathbf{V}, \boldsymbol{\pi}(V_i) \in \Sigma^{|\mathbf{\Pi}(V_i)|}\}, \{\mathbf{Pr}[U_i \mid \boldsymbol{\pi}(U_i)] : U_i \in \mathbf{U}, \boldsymbol{\pi}(U_i) \in \Sigma^{|\mathbf{\Pi}(U_i)|}\}\rangle$*, where (i)* $\mathbf{V}$ *and* $\mathbf{U}$ *are the sets of observable and unobservable variables respectively, (ii)* G *is a directed acyclic graph on* $\mathbf{V} \cup \mathbf{U}$*, and (iii)* $\mathbf{Pr}[V_i \mid \boldsymbol{\pi}(V_i)]$ *and* $\mathbf{Pr}[U_i \mid \boldsymbol{\pi}(U_i)]$ *are the conditional probability distributions of* V_i *and* U_i *resp. given that its parents* $\mathbf{\Pi}(V_i)$ *and* $\mathbf{\Pi}(U_i)$ *resp. take the values* $\boldsymbol{\pi}(V_i)$ *and* $\boldsymbol{\pi}(U_i)$*) resp.*

A CBN $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, G, \{\mathbf{Pr}[V_i \mid \boldsymbol{\pi}(V_i)]\}, \{\mathbf{Pr}[U_i \mid \boldsymbol{\pi}(U_i)]\}\rangle$ *defines a unique interventional distribution* $P_{\mathcal{M}}[\mathbf{V} \mid do(\mathbf{x})]$ *for every subset* $\mathbf{X} \subseteq \mathbf{V}$ *(including* $\mathbf{X} = \emptyset$*) and assignment* $\mathbf{x} \in \Sigma^{|\mathbf{X}|}$*, as follows. For all* $\mathbf{v} \in \Sigma^{|\mathbf{V}|}$*:*

$$P_{\mathcal{M}}[\mathbf{v} \mid do(\mathbf{x})] = \begin{cases} \sum_{\mathbf{u}} \prod_{V_i \in \mathbf{V} \setminus \mathbf{X}} \mathbf{Pr}[v_i \mid \boldsymbol{\pi}(V_i)] \cdot \prod_{U_i \in \mathbf{U}} \mathbf{Pr}[u_i \mid \boldsymbol{\pi}(U_i)] & \text{if } \mathbf{v} \text{ is consistent with } \mathbf{x} \\ 0 & \text{otherwise.} \end{cases}$$

We say that G *is the* causal graph *corresponding to the* CBN $\mathcal{M}$*.*

Another equivalent way to define a CBN is by specifying the set of interventional distributions $P_{\mathcal{M}}[\mathbf{V} \mid do(\mathbf{x})]$ for all subsets $\mathbf{X}$ and assignments $\mathbf{x}$. To connect to the preceding definition, we require that each $P_{\mathcal{M}}[\mathbf{V} \mid do(\mathbf{x})]$ is defined by the Bayesian network described by $G_{\overline{\mathbf{X}}}$ with the conditional probability distributions obtained by setting the variables in $\mathbf{X}$ to the constants $\mathbf{x}$.

It is standard in the causality literature to work with causal graphs of a particular structure:

Definition 2.9 (Semi-Markovian causal graph and Semi-Markovian Bayesian network)**.** *A* semi-Markovian causal graph *(*SMCG*) G is a directed acyclic graph on* $\mathbf{V} \cup \mathbf{U}$ *where every unobservable variable is a root node and has exactly two children, both observable. A* semi-Markovian Bayesian network (SMBN) *is a causal Bayesian network where the causal graph is semi-Markovian.*

There exists a known reduction (described formally in Appendix B) from general causal Bayesian networks to semi-Markovian Bayesian networks that preserves all the properties we use in our analysis, so that henceforth, we will restrict only to SMBNs.

In SMCGs, the divergent edges $V_i \leftarrow U_k \rightarrow V_j$ are usually represented by *bi-directed edges* $V_i \leftrightarrow V_j$. A bi-directed edge between two observable variables implicitly represents the presence of an unobservable parent.

Definition 2.10 (c-component)**.** *For a given* SMCG 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 bi-directed edges.*

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. We use the notation $C(\mathbf{V}) = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_k\}$ to denote the partition of $\mathbf{V}$ into the c-components of G, where each $\mathbf{S}_i \subseteq \mathbf{V}$ is a c-component of G.

Also, for $\mathbf{X} \subseteq \mathbf{V}$, the induced subgraph $G[\mathbf{X}]$ is the subgraph obtained by removing the vertices $\mathbf{V} \setminus \mathbf{X}$ and their corresponding edges from G. We use the notation $C(\mathbf{X}) = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_k\}$ to denote the set of all c-components of $G[\mathbf{X}]$, that is each $\mathbf{S}_i \subseteq \mathbf{X}$ is a c-component of $G[\mathbf{X}]$. The next two lemmas capture the factorizations of distributions in SMBN.

Lemma 2.11. *Let* $\mathcal{M}$ *be a given* SMBN *with respect to the* SMCG G*. For any set* $\mathbf{S} \subseteq \mathbf{V}$*, and a subset* $\mathbf{D}$ *such that* $(\mathbf{V} \setminus \mathbf{S}) \supseteq \mathbf{D} \supseteq \mathbf{Pa}(\mathbf{S})$*, and for any assignment* $\mathbf{s}, \mathbf{d}$*,*

$$P_{\mathcal{M}}[\mathbf{s} \mid do(\mathbf{d})] = P_{\mathcal{M}}[\mathbf{s} \mid do(\mathbf{pa}(\mathbf{S}))]$$

where $\mathbf{pa}(\mathbf{S})$ *is consistent with the assignment* $\mathbf{d}$*.*

Proof. When the parents of $\mathbf{S}$, $\mathbf{Pa}(\mathbf{S})$, are targeted for intervention, the distribution on $\mathbf{S}$ remains the same irrespective of whether the other vertices in $(\mathbf{V} \setminus \mathbf{S})$ are intervened or not. □

Lemma 2.12 (c-component factorization, [TP02])**.** *Given a* SMBN $\mathcal{M}$ *with respect to the causal graph* G *and a subset* $\mathbf{X} \subseteq \mathbf{V}$*, let* $C(\mathbf{V} \setminus \mathbf{X}) = \{\mathbf{S}_1, \ldots, \mathbf{S}_k\}$*. For any given assignment* $\mathbf{v}$*,*

$$P_{\mathcal{M}}[\mathbf{v} \setminus \mathbf{x} \mid do(\mathbf{x})] = \prod_i P_{\mathcal{M}}[\mathbf{s}_i \mid do(\mathbf{v} \setminus \mathbf{s}_i)].$$

For a given SMCG G, the in-degree and out-degree of an observable vertex $V_i \in \mathbf{V}$ denote the number of observable parents and observable children of V_i in G respectively. The maximum in-degree of a SMCG G is the maximum in-degree over all the observable vertices. The maximum degree of a SMCG G is the maximum of the sum of the in-degree and out-degree over all the observable vertices.

Definition 2.13 (Graphs with *bounded in-degree* and *bounded c-component*)**.** $\mathcal{G}_{d,\ell}$ *denotes the class of* SMCG*s with maximum in-degree at most* d *and the size of the largest c-component at most* ℓ*.*

2.2 Problem Definitions

Here we define the testing and learning problems considered in the paper. Let $\mathcal{M}$ and $\mathcal{N}$ be two SMBNs. We say that $\mathcal{M} = \mathcal{N}$, if

$$P_{\mathcal{M}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})] = P_{\mathcal{N}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})] \qquad \forall \mathbf{T} \subseteq \mathbf{V}, \mathbf{t} \in \Sigma^{|\mathbf{T}|}.$$

And we say that $\Delta(\mathcal{M}, \mathcal{N}) > \varepsilon$, if there exists $\mathbf{T} \subseteq \mathbf{V}$ and $\mathbf{t} \in \Sigma^{|\mathbf{T}|}$ such that

$$\delta_{TV}(P_{\mathcal{M}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})], P_{\mathcal{N}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})]) > \varepsilon.$$

Definition 2.14 (Causal Goodness-of-fit Testing ($\mathsf{CGFT}(G, \mathcal{M}, \varepsilon)$))**.** *Given a* SMCG G*, a (known)* SMBN $\mathcal{M}$ *on* G*, and* $\varepsilon > 0$*. Let* $\mathcal{X}$ *denote an unknown* SMBN *on* G*. The objective of* $\mathsf{CGFT}(G, \mathcal{M}, \varepsilon)$ *is to distinguish between* $\mathcal{X} = \mathcal{M}$ *versus* $\Delta(\mathcal{X}, \mathcal{M}) > \varepsilon$ *with probability at least 2/3, by performing interventions and taking samples from the resulting interventional distributions of* $\mathcal{X}$*.*

Definition 2.15 (Causal Two-sample Testing ($\mathsf{C2ST}(G,\varepsilon)$)). *Given a* SMCG G*, and* $\varepsilon > 0$*. Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be two unknown* SMBN*s on* G*. The objective of* $\mathsf{C2ST}(G,\varepsilon)$ *is to distinguish between* $\mathcal{X} = \mathcal{Y}$ *versus* $\Delta(\mathcal{X},\mathcal{Y}) > \varepsilon$ *with probability at least 2/3, by performing interventions and taking samples from the resulting interventional distributions of* $\mathcal{X}$ *and* $\mathcal{Y}$*.*

Definition 2.16 (Learning SMBNs ($\mathsf{CL}(G,\varepsilon)$)). *Given a* SMCG G *and* $\varepsilon > 0$*. Let* $\mathcal{X}$ *be an unknown* SMBN *on* G*. The objective of* $\mathsf{CL}(G,\varepsilon)$ *is to perform interventions and taking samples from the resulting interventional distributions of* $\mathcal{X}$*, and return an oracle that for any* $\mathbf{T} \subseteq \mathbf{V}$ *and* $\mathbf{t} \in \Sigma^{|\mathbf{T}|}$ *returns an estimated interventional distribution* $P_{ES}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})]$ *such that*

$$\delta_{TV}([P_{\mathcal{X}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})], P_{ES}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})]) < \varepsilon.$$

We emphasize that in all three problems, the causal graph G is known explicitly in advance.

3 Testing and Learning Algorithms for SMBNs

Before we discuss our algorithms, we begin by defining *covering intervention sets.*

Definition 3.1. *A set of interventions* $\mathbf{I}$ *is a* covering intervention set *if for every subset* $\mathbf{S}$ *of every c-component, and every assignment* $\mathbf{pa}(\mathbf{S}) \in \Sigma^{|\mathbf{Pa}(\mathbf{S})|}$ *there exists an* $I \in \mathbf{I}$ *such that,*

- *No node in* $\mathbf{S}$ *is intervened in* I*.*
- *Every node in* $\mathbf{Pa}(S)$ *is intervened.*
- I *restricted to* $\mathbf{Pa}(S)$ *has the assignment* $\mathbf{pa}(S)$*.*

Our algorithms comprise of two key arguments.

- A procedure to compute a covering intervention set $\mathbf{I}$ of *small size*.
- A sub-additivity result for CBNs that allows us to localize the distances: where we show that two CBNs are far implies there exist a marginal distribution of some intervention in $\mathbf{I}$ such that the marginals are far.

These two results are formalized in Section 4.1, and Section 4.2 respectively.

3.1 Testing

Our main testing result is the following upper bound for testing of causal models.

Theorem 3.2 (Algorithm for $\mathsf{C2ST}(G,\varepsilon)$). *Let* G *be a* SMCG $\in \mathcal{G}_{d,\ell}$ *with* n *vertices. Let the variables take values over a set* Σ *of size* K*. Then, there is an algorithm to solve* $\mathsf{C2ST}(G,\varepsilon)$*, that makes* $O(K^{\ell d}(3d)^{\ell}\log n)$ *interventions to each of the unknown* SMBN*s* $\mathcal{X}$ *and* $\mathcal{Y}$*, taking* $\tilde{O}(K^{\ell(d+7/4)}n\varepsilon^{-2})$ *samples per intervention, in time* $\tilde{O}(2^{\ell}K^{\ell(2d+7/4)}n^2\varepsilon^{-2})$*.*

When the maximum degree (in-degree plus out-degree) of G *is bounded by* d*, then our algorithm uses* $O(K^{\ell d}(3d)^{\ell}\ell d^2\log K)$ *interventions with the same sample complexity and running time as above.*

This result gives Theorem 1.1 as a corollary, since two sample tests are harder than one sample tests.

Proof of Theorem 3.2. Our algorithm is described in Algorithm 1.

The algorithm starts with a covering intervention set $\mathbf{I}$. Lemma 4.1 gives an $\mathbf{I}$ with $O(K^{\ell d}(3d)^{\ell}(\log n + \ell d \log K))$ interventions. When the maximum degree is bounded by d, then Lemma 4.3 gives an $\mathbf{I}$ of size $O(K^{\ell d}(3d)^{\ell}\ell d^2 \log K)$. Moreover, by the remarks following Lemmas 4.1 and 4.3, $\mathbf{I}$ can be found in $\tilde{O}(n)$ time.

Algorithm 1: Algorithm for $\mathsf{C2ST}(G, \varepsilon)$

$\mathbf{I}$: Covering intervention set

1. Under each intervention $I \in \mathbf{I}$:
   (a) Obtain $\tilde{O}(K^{\ell(d+7/4)}n\varepsilon^{-2})$ samples from the interventional distribution of I in both models $\mathcal{X}$ and $\mathcal{Y}$.
   (b) For any subset $\mathbf{S}$ of a c-component of G, if I does not set $\mathbf{S}$ but sets $\mathbf{Pa}(\mathbf{S})$ to $\mathbf{pa}(\mathbf{S})$, then using Lemma 2.2, Lemma 2.11 and the obtained samples, test (with error probability at most $1/(3K^{\ell d}2^{\ell}n)$):

   $$P_{\mathcal{X}}[\mathbf{S} \mid do(\mathbf{pa}(\mathbf{S}))] = P_{\mathcal{Y}}[\mathbf{S} \mid do(\mathbf{pa}(\mathbf{S}))] \text{ versus } H^2\left(\begin{array}{c} P_{\mathcal{X}}[\mathbf{S} \mid do(\mathbf{pa}(\mathbf{S}))], \\ P_{\mathcal{Y}}[\mathbf{S} \mid do(\mathbf{pa}(\mathbf{S}))] \end{array}\right) \geqslant \frac{\varepsilon^2}{2K^{\ell(d+1)}n}$$

   Output "$\Delta(\mathcal{X}, \mathcal{Y}) > \varepsilon$" if the latter.
2. Output "$\mathcal{X} = \mathcal{Y}$".

We will now analyze the performance of our algorithm.

Number of interventions, time, and sample requirements. The number of interventions is the size of $\mathbf{I}$, bounded from Lemma 4.1 or Lemma 4.3. The number of samples per intervention is given in the algorithm. The algorithm performs $n2^{\ell}K^{\ell d}$ sub-tests. And for each such sub-test, the algorithm's running time is quasi-linear in the sample complexity (Lemma 2.2), therefore taking a total time of $\tilde{O}(2^{\ell}K^{\ell(2d+7/4)}n^2\varepsilon^{-2})$.

Correctness. In Theorem 4.5, we show that when $\Delta(\mathcal{X}, \mathcal{Y}) > \varepsilon$, there exists a subset $\mathbf{S}$ of some c-component, and an $I \in \mathbf{I}$ that does not intervene any node in $\mathbf{S}$ but intervenes $\mathbf{Pa}(\mathbf{S})$ with some assignment $\mathbf{pa}(\mathbf{s})$ such that

$$H^2(P_{\mathcal{X}}[\mathbf{S} \mid do(\mathbf{pa}(\mathbf{S}))], P_{\mathcal{Y}}[\mathbf{S} \mid do(\mathbf{pa}(\mathbf{S}))]) > \varepsilon^2/(2K^{\ell(d+1)}n).$$

This structural result is the key to our algorithm. This together with Lemma 2.1 proves that $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ are far in terms of the total variation distance. To bound the error probability, note that the number of total sub-tests we run is bounded by $K^{\ell d}n2^{\ell}$, and the error probability for each subset is at most $1/(3K^{\ell d}2^{\ell}n)$, by the union bound, we will have an error of at most $1/3$ over the entire algorithm. □

In some cases, the underlying SMCG might not be known. We will now consider the problem of two sample testing, where $\mathcal{X}$ and $\mathcal{Y}$ are still on the same common SMCG G, but G is unknown. We now show an algorithm that uses the same number of interventions and samples as Theorem 3.2 for the known G case, however requiring $O(n^{\ell+1}K^{\ell(2d+7/4)}\varepsilon^{-2})$ time.

Theorem 3.3 (Algorithm for $\mathsf{C2ST}(G, \varepsilon)$ – Unknown graph)**.** *Consider the same set-up as Theorem 3.2, except that the* SMCG $G \in \mathcal{G}_{d,\ell}$ *is unknown. Then, there is an algorithm to this problem, that makes*

$O(K^{\ell d}(3d)^\ell \log n)$ *interventions to* $\mathcal{X}$ *and* $\mathcal{Y}$*, taking* $\tilde{O}(K^{\ell(d+7/4)} n\varepsilon^{-2})$ *samples per intervention, in time* $\tilde{O}(n^\ell K^{\ell(2d+7/4)} n\varepsilon^{-2})$.

Proof. We first use Lemma 4.1 and obtain a set of interventions $\mathbf{I}$, such that $\mathbf{I}$ is a covering set with error probability at most $1/6$. Note that Lemma 4.1 holds even when the underlying graph G is unknown.

Algorithm 2: Algorithm for $\mathsf{C2ST}(G, \varepsilon)$ – Unknown graph

$\mathbf{I}$: Covering intervention set

1. Under each intervention $I = \mathbf{Pr}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})] \in \mathbf{I}$:
   (a) Obtain $\tilde{O}(K^{\ell(d+7/4)} n\varepsilon^{-2})$ samples from the interventional distribution of I in both models $\mathcal{X}$ and $\mathcal{Y}$.
   (b) For each subset $\mathbf{S} \subseteq \mathbf{V} \setminus \mathbf{T}$ of size $\leqslant \ell$, using Lemma 2.2, Lemma 2.11 and the obtained samples, test (with error probability at most $1/(6K^{\ell d} 2^\ell n)$):

$$P_{\mathcal{X}}[\mathbf{S} \mid do(\mathbf{t})] = P_{\mathcal{Y}}[\mathbf{S} \mid do(\mathbf{t})] \quad \text{versus} \quad H^2 \begin{pmatrix} P_{\mathcal{X}}[\mathbf{S} \mid do(\mathbf{t})], \\ P_{\mathcal{Y}}[\mathbf{S} \mid do(\mathbf{t})] \end{pmatrix} \geqslant \frac{\varepsilon^2}{2K^{\ell(d+1)} n}$$

   Output "$\Delta(\mathcal{X}, \mathcal{Y}) > \varepsilon$" if the latter.
2. Output "$\mathcal{X} = \mathcal{Y}$".

For each intervention, we go over all subsets S of size $\leqslant \ell$. Therefore we perform at most $\binom{n}{\leqslant \ell} = O(n^\ell)$ sub-tests for an intervention. For each sub-test, the algorithm's running time is quasi-linear in the sample complexity (Lemma 2.2), therefore taking a total time of $O(n^\ell K^{\ell(2d+7/4)} n\varepsilon^{-2})$. The number of interventions follow from Lemma 4.1 and the number of samples follow from the algorithm.

Correctness. As in the proof of Theorem 3.2, we use Theorem 4.5 to show that when $\Delta(\mathcal{X}, \mathcal{Y}) > \varepsilon$, then there exists a subset $\mathbf{S}$ of some c-component and an $I \in \mathbf{I}$ that does not intervene any node in $\mathbf{S}$ but intervenes $\mathbf{Pa}(\mathbf{S})$ with some assignment $\mathbf{pa}(\mathbf{s})$ such that

$$H^2(P_{\mathcal{X}}[\mathbf{S} \mid do(\mathbf{pa}(\mathbf{S}))], P_{\mathcal{Y}}[\mathbf{S} \mid do(\mathbf{pa}(\mathbf{S}))]) > \varepsilon^2/(2K^{\ell(d+1)} n).$$

This together with Lemma 4.5 proves that $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ are far in terms of the total variation distance. Since the error probability of each sub-test is bounded by at most $1/(6K^{\ell d} 2^\ell n)$ and the error probability of $\mathbf{I}$ being a covering intervention set is at most $1/6$, by union bound, we will have an error of at most $1/3$ over the entire algorithm. □

3.2 Learning

Our next result is on learning SMBNs over a known causal graph. Our algorithm is improper, meaning that it does not output a causal model in the form of an SMBN, but rather outputs an oracle which succinctly encodes all the interventional distributions. See Definition 2.16 for a rigorous formulation of the problem.

Theorem 3.4 (Algorithm for $\mathsf{CL}(G, \varepsilon)$)**.** *For any given* SMCG $G \in \mathcal{G}_{d,\ell}$ *with* n *vertices and a parameter* $\varepsilon > 0$*, there exists an algorithm that takes as input an unknown* SMBN $\mathcal{X}$ *over* G*, that performs* $O(K^{\ell d}(3d)^\ell \log n)$ *interventions to* $\mathcal{X}$*, taking* $\tilde{O}(K^{\ell(2d+3)} n^2 \varepsilon^{-4})$ *samples per intervention, that runs in*

time $\tilde{O}\left(2^{\ell}K^{\ell(3d+3)}n^3\varepsilon^{-4}\right)$*, and that with probability at least* $2/3$*, outputs an oracle* $\mathcal{N}$ *with the following behavior. Given as input any* $\mathbf{T} \subseteq \mathbf{V}$ *and assignment* $\mathbf{t} \in \Sigma^{|\mathbf{T}|}$*,* $\mathcal{N}$ *outputs an interventional distribution* $P_{\mathcal{N}}[\mathbf{V} \setminus \mathbf{T}|do(\mathbf{t})]$ *such that:*

$$\delta_{TV}(P_{\mathcal{X}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})], P_{\mathcal{N}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})]) < \varepsilon$$

When the maximum degree (in-degree plus out-degree) of G *is bounded by* d*, then our algorithm uses* $O(K^{\ell d}(3d)^{\ell}\ell d^2 \log K)$ *interventions with the same sample complexity and running time as above.*

Algorithm 3: Algorithm for $\mathsf{CL}(G, \varepsilon)$

I: Covering intervention set

1. Under each intervention $I \in \mathbf{I}$:
   (a) Obtain $\tilde{O}(n^2K^{\ell(2d+3)}\varepsilon^{-4})$ samples from the interventional distribution of I in $\mathcal{X}$.
   (b) For each subset $\mathbf{S}$ of a c-component, if I does not set $\mathbf{S}$ but sets $\mathbf{Pa}(\mathbf{S})$ to $\mathbf{pa}(\mathbf{S})$, use Lemma 2.3, Lemma 2.11 and the obtained samples to learn:

   $$P_{\mathcal{N}}[\mathbf{S} \mid do(\mathbf{pa}(\mathbf{S}))] \text{ such that } H^2(P_{\mathcal{N}}[\mathbf{S} \mid do(\mathbf{pa}(\mathbf{S}))], P_{\mathcal{X}}[\mathbf{S} \mid do(\mathbf{pa}(\mathbf{S}))]) \leqslant \frac{\varepsilon^2}{2K^{\ell(d+1)}n}$$

   with probability of error at most $1/(3K^{\ell d}2^{\ell}n)$.
2. Return the following oracle $\mathcal{N}$ that takes as input: $\mathbf{T} \subseteq \mathbf{V}$ and $\mathbf{t} \in \Sigma^{|\mathbf{T}|}$

   (i) Let $C(\mathbf{V} \setminus \mathbf{T}) = \{\mathbf{S}_1, \ldots, \mathbf{S}_p\}$.
   (ii) Output the distribution $P_{\mathcal{N}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})]$ where for any assignment $\mathbf{v} \setminus \mathbf{t}$:

   $$P_{\mathcal{N}}[\mathbf{v} \setminus \mathbf{t} \mid do(\mathbf{t})] = \prod_{i=1}^{p} P_{\mathcal{N}}[\mathbf{s}_i \mid do(\mathbf{v} \setminus \mathbf{s}_i)]$$

The covering intervention set used in the algorithm above is as defined in Definition 3.1.

Number of interventions, time, and sample requirements. The number of interventions is obtained using the bound on the size of the covering intervention set from Lemma 4.1. When the maximum degree is bounded, we can use Lemma 4.3. The number of samples per intervention is obtained from Lemma 2.3. Since the algorithm learns at most $nK^{\ell d}2^{\ell}$ interventions (subroutines), and each subroutine takes time linear in the sample size, the time complexity follows.

Correctness. For any given $\mathbf{T}$, $do(\mathbf{t})$, let $C(\mathbf{V} \setminus \mathbf{T}) = \{\mathbf{S}_1, \ldots, \mathbf{S}_p\}$. Lemma 2.12 justifies that

$$P_{\mathcal{N}}[\mathbf{v} \setminus \mathbf{t} \mid do(\mathbf{t})] = \prod_{i} P_{\mathcal{N}}[\mathbf{s}_i \mid do(\mathbf{v} \setminus \mathbf{s}_i)].$$

Similar to the proof of Theorem 3.2, using Theorem 4.5 and Lemma 2.1, we get:

$$\begin{aligned} & H^2(P_{\mathcal{N}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})], P_{\mathcal{X}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})]) < \varepsilon^2/2 \\ \implies & \delta_{TV}(P_{\mathcal{N}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})], P_{\mathcal{X}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})]) < \varepsilon. \end{aligned}$$

4 Main Ingredients of the Analysis

4.1 Covering Intervention Sets

Lemma 4.1 (Counting Lemma: bounded in-degree)**.** *Let $G \in \mathcal{G}_{d,\ell}$ be a* SMCG *with n vertices and Σ be an alphabet set of size K. Then, there is a randomized algorithm that outputs a set $\mathbf{I}$ of size $O(K^{\ell d}(3d)^{\ell}(\log n + \ell d \log K + \log(1/\delta)))$. such that, with probability at least $1-\delta$, $\mathbf{I}$ is a covering intervention set.*

Proof. Let $t = K^{\ell d}(3d)^{\ell}(\log n + 2\ell d \log K + \log(1/\delta))$. The interventions in $\mathbf{I}$ are chosen by the following procedure: For each $j \in [t]$ and for each $V_i \in V$, V_i is observed in I_j with probability $1/(d+1)$ and otherwise, V_i is intervened with the assignment chosen uniformly from Σ. Let $V_i = *$ denotes that V_i is not intervened. Consider a fixed c-component $\mathbf{C}$, a fixed subset $\mathbf{S} \subseteq \mathbf{C}$, a fixed assignment $\mathbf{pa}(\mathbf{S}) \in \Sigma^{|\mathbf{Pa}(\mathbf{S})|}$ and a fixed $j \in [t]$. Now,

$$\begin{aligned}\mathbf{Pr}[I_j(\mathbf{S}) = *^{|\mathbf{S}|} \wedge I_j(\mathbf{Pa}(\mathbf{S})) = \mathbf{pa}(\mathbf{S})] &= \left(\frac{1}{d+1}\right)^{|\mathbf{S}|} \cdot \left(\frac{d}{K(d+1)}\right)^{|\mathbf{Pa}(\mathbf{S})|} \\ &\geqslant (d+1)^{-\ell} K^{-\ell d} e^{-\ell} \qquad [\text{Since } |\mathbf{Pa}(\mathbf{S})| \leqslant \ell d \text{ and } |\mathbf{S}| \leqslant \ell] \\ &\geqslant (3d)^{-\ell} K^{-\ell d}.\end{aligned}$$

This implies that

$$\mathbf{Pr}[\forall j \in [t], (I_j(\mathbf{S}) \neq *^{|\mathbf{S}|} \vee I_j(\mathbf{Pa}(\mathbf{S})) \neq \mathbf{pa}(\mathbf{S}))] \leqslant \left(1 - (3d)^{-\ell} K^{-\ell d}\right)^t \leqslant \frac{\delta}{n} K^{-2\ell d}.$$

Hence,

$$\begin{aligned}&\mathbf{Pr}[\exists\, C\text{-component } \mathbf{C}, \exists \mathbf{S} \subseteq C, \exists\, \mathbf{pa}(\mathbf{S}) \in \Sigma^{|\mathbf{Pa}(\mathbf{S})|}, \forall j \in [t], (I_j(\mathbf{S}) \neq *^{|\mathbf{S}|} \vee I_j(\mathbf{Pa}(\mathbf{S})) \neq \mathbf{pa}(\mathbf{S}))] \\ &\leqslant n 2^{\ell} K^{\ell d} \cdot \frac{\delta}{n} K^{-2\ell d} \leqslant \delta\end{aligned}$$

by the union bound. □

Remark 4.2. *The above proof can be made deterministic by using explicit deterministic constructions of almost ℓd-wise independent random variables [AGHP92, EGL+92].*

Lemma 4.3 (Counting Lemma: bounded total degree)**.** *Let $G \in \mathcal{G}_{d,\ell}$ be an* SMCG *with n vertices, whose variables take values in Σ with $|\Sigma| = K$, and whose maximum degree is bounded by d. Then, there exists covering intervention set $\mathbf{I}$ of size $O(K^{\ell d}(3d)^{\ell}\ell d^2 \log K)$.*

Proof. Let $t = K^{\ell d}(3d)^{\ell}(\ell d^2 + \ell d \log K + 2)$. The interventions in $\mathbf{I}$ are chosen by the following procedure: For each $j \in [t]$ and for each $V_i \in V$, V_i is observed in I_j with probability $1/(d+1)$ and otherwise, V_i is intervened with the assignment chosen uniformly from the set Σ. Let $V_i = *$ denotes that V_i is observed (not intervened).

For a fixed set $\mathbf{S}$ that is a subset of a c-component and a fixed assignment $\mathbf{pa}(\mathbf{S}) \in \Sigma^{|\mathbf{Pa}(\mathbf{S})|}$, let $A_{\mathbf{S},\mathbf{pa}(\mathbf{S})}$ be the event: $\forall j \in [t], (I_j(\mathbf{S}) \neq *^{|\mathbf{S}|} \vee I_j(\mathbf{Pa}(\mathbf{S})) \neq \mathbf{pa}(\mathbf{S}))$. Similar to the proof of Lemma 4.1, for any fixed $\mathbf{S}$ and $\mathbf{pa}(\mathbf{S})$: $\mathbf{Pr}[A_{\mathbf{S},\mathbf{pa}(\mathbf{S})}] \leqslant 1/(42^{\ell d^2} K^{\ell d})$.

Now, note that $A_{\mathbf{S},\mathbf{pa}(\mathbf{S})}$ and $A_{\mathbf{T},\mathbf{pa}(\mathbf{T})}$ are independent if $\mathbf{Pa}(\mathbf{S})$ and $\mathbf{Pa}(\mathbf{T})$ are disjoint. For a fixed $\mathbf{S}$, the number of subsets $\mathbf{T}$ such that $\mathbf{Pa}(\mathbf{S}) \cap \mathbf{Pa}(\mathbf{T}) \neq \emptyset$ is at most $2^{\ell d^2}$ (since, the number of children of the parents of S is at most ℓd^2). Therefore, for a fixed $\mathbf{S}$ and $\mathbf{pa}(\mathbf{S})$, $A_{\mathbf{S},\mathbf{pa}(\mathbf{S})}$ is independent of all $A_{\mathbf{T},\mathbf{pa}(\mathbf{T})}$'s

except for at most $2^{\ell d^2} K^{\ell d}$ many of them (taking into account the number of possible assignments $\mathbf{pa}(\mathbf{T})$). Hence, the Lovász Local Lemma [AS04, Chapter 5] guarantees that there exists a set of t interventions such that $\neg A_{\mathbf{S},\mathbf{pa}(\mathbf{S})}$ for all $\mathbf{S}$ and $\mathbf{pa}(\mathbf{S})$. □

Remark 4.4 (Explicitness)**.** *Although Lemma 4.3 only asserts the existence of a covering intervention, its proof can be turned into a linear time algorithm using the constructive proofs of the Lovász Local Lemma [Mos09, MT10].*

4.2 Subadditivity Theorem for SMBNs

The next theorem states that if two causal models are "far", then they must be "far" under some "local" intervention.

Theorem 4.5. *Let $\mathcal{M}$ and $\mathcal{N}$ be two* SMBN*s defined on a* known and common SMCG $G \in \mathcal{G}_{d,\ell}$*. Let $\mathbf{V}$ be the vertices of G. For a given intervention $do(\mathbf{t})$, let $\mathbf{V} \setminus \mathbf{T}$ partition into $\mathcal{C} = \{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_p\}$, the c-components with respect to the induced graph $G[\mathbf{V} \setminus \mathbf{T}]$. Suppose*

$$H^2(P_{\mathcal{M}}[\mathbf{C}_j \mid do(\mathbf{pa}(\mathbf{C}_j))], P_{\mathcal{N}}[\mathbf{C}_j \mid do(\mathbf{pa}(\mathbf{C}_j))]) \leqslant \gamma \qquad \forall j \in [p], \forall\, \mathbf{pa}(\mathbf{C}_j) \in \Sigma^{|\mathbf{Pa}(\mathbf{C}_j)|}. \tag{2}$$

Then

$$H^2\left(P_{\mathcal{M}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})], P_{\mathcal{N}}[\mathbf{V} \setminus \mathbf{T} \mid do(\mathbf{t})]\right) \leqslant \varepsilon \qquad \forall \mathbf{t} \in \Sigma^{|\mathbf{T}|} \tag{3}$$

where $\varepsilon = \gamma |\Sigma|^{\ell(d+1)} n$.

Proof. Let $\mathbf{W} = \mathbf{V} \setminus \mathbf{T} = \{W_1, \ldots, W_r\}$, where the indices are arranged in a topological ordering. Here we focus only on distributions on $\mathbf{W}$ after the intervention $do(\mathbf{t})$. That is, our focus is restricted to the graph $G_{\overline{\mathbf{T}}}$, the intervention $do(\mathbf{t})$ and the vertices $\mathbf{W} = \mathbf{V} \setminus \mathbf{T}$. We know that

$$\begin{aligned} H^2\left(P_{\mathcal{M}}[\mathbf{W} \mid do(\mathbf{t})], P_{\mathcal{N}}[\mathbf{W} \mid do(\mathbf{t})]\right) &= 1 - \sum_{\mathbf{w}} \sqrt{P_{\mathcal{M}}[\mathbf{w} \mid do(\mathbf{t})], P_{\mathcal{N}}[\mathbf{w} \mid do(\mathbf{t})]} \\ &= 1 - BC\left(P_{\mathcal{M}}[\mathbf{W} \mid do(\mathbf{t})], P_{\mathcal{N}}[\mathbf{W} \mid do(\mathbf{t})]\right) \end{aligned} \tag{4}$$

where $BC(P_{\mathcal{M}}[\mathbf{W} \mid do(\mathbf{t})], P_{\mathcal{N}}[\mathbf{W} \mid do(\mathbf{t})])$ is the Bhattacharya coefficient of $P_{\mathcal{M}}[\mathbf{W} \mid do(\mathbf{t})]$ and $P_{\mathcal{N}}[\mathbf{W} \mid do(\mathbf{t})]$ (see (9)).
For each $j \in [p]$, identify the vertices in $\mathbf{C}_j$ as $\{W_{n_{j,1}}, \ldots, W_{n_{j,s_j}}\}$ where $s_j = |\mathbf{C}_j|$ and $n_{j,1} < \cdots < n_{j,s_j}$. Using Lemma 2.12, we express the distributions in terms of the product $\prod_{j=1}^{p} \mathbf{Pr}[\mathbf{c}_j \mid do(\mathbf{w} \setminus \mathbf{c}_j)]$ [TP02],

$$\begin{aligned} & BC(P_{\mathcal{M}}[\mathbf{W} \mid do(\mathbf{t})], P_{\mathcal{N}}[\mathbf{W} \mid do(\mathbf{t})]) \\ &= \sum_{\mathbf{w}} \sqrt{P_{\mathcal{M}}[\mathbf{w} \mid do(\mathbf{t})] \; P_{\mathcal{N}}[\mathbf{w} \mid do(\mathbf{t})]} \\ &= \sum_{\mathbf{w}} \sqrt{\prod_{j=1}^{p} P_{\mathcal{M}}[\mathbf{c}_j \mid do(\mathbf{w} \setminus \mathbf{c}_j)] \; P_{\mathcal{N}}[\mathbf{c}_j \mid do(\mathbf{w} \setminus \mathbf{c}_j)]} \\ &= \sum_{\mathbf{w}} \sqrt{\prod_{j=1}^{p} \prod_{i=1}^{s_j} P_{\mathcal{M}}[w_{n_{j,i}} \mid w_{n_{j,1}}, \ldots, w_{n_{j,i-1}}, do(\mathbf{w} \setminus \mathbf{c}_j)] \; P_{\mathcal{N}}[w_{n_{j,i}} \mid w_{n_{j,1}}, \ldots, w_{n_{j,i-1}}, do(\mathbf{w} \setminus \mathbf{c}_j)]} \\ &= \sum_{\mathbf{w}} \sqrt{\prod_{j=1}^{p} \prod_{i=1}^{s_j} P_{\mathcal{M}}[w_{n_{j,i}} \mid w_{n_{j,1}}, \ldots, w_{n_{j,i-1}}, do(\mathbf{pa}(W_{n_{j,1}}, \ldots, w_{n_{j,i-1}}))] \; P_{\mathcal{N}}[w_{n_{j,i}} \mid w_{n_{j,1}}, \ldots, w_{n_{j,i-1}}, do(\mathbf{pa}(W_{n_{j,1}}, \ldots, w_{n_{j,i-1}}))]} \end{aligned}$$

(using Lemma C.1).

For $i \in [s_j]$, let

$$dep(n_{j,i}) := \{W_{n_{j,1}}, \ldots, W_{n_{j,i}}\} \cup (\mathbf{Pa}(\{W_{n_{j,1}}, \ldots, W_{n_{j,i}}\}) \setminus \mathbf{T}).$$

For $j \in [p], i \in [s_j]$, let $X_{n_{j,i}} : \Sigma^{|dep(n_{j,i})|} \to [0,1]$ be

$$X_{n_{j,i}}(\mathbf{w}_{dep(n_{j,i})}) := \sqrt{\frac{P_{\mathcal{M}}[w_{n_{j,i}} \mid w_{n_{j,1}}, \ldots, w_{n_{j,i-1}}, do(\mathbf{pa}(W_{n_{j,1}}, \ldots, w_{n_{j,i-1}}))]}{P_{\mathcal{N}}[w_{n_{j,i}} \mid w_{n_{j,1}}, \ldots, w_{n_{j,i-1}}, do(\mathbf{pa}(W_{n_{j,1}}, \ldots, w_{n_{j,i-1}}))]}}.$$

Recall that indices of $\mathbf{W}$ follow a topological ordering. Using this topological ordering and plugging in the expression above, we obtain

$$BC(P_{\mathcal{M}}[\mathbf{W} \mid do(\mathbf{t})], P_{\mathcal{N}}[\mathbf{W} \mid do(\mathbf{t})]) = \sum_{w_1} X_1(\mathbf{w}_{dep(1)}) \sum_{w_2} X_2(\mathbf{w}_{dep(2)}) \ldots \sum_{w_r} X_r(\mathbf{w}_{dep(r)})$$

where $r = |\mathbf{W}|$. In order to prove the theorem, it will suffice to prove that this expression is at least $1 - \varepsilon$, whenever (2) holds. To prove this, we will take the following path, which is essentially an induction on r. For $j \in [p]$, let $b_j = 1$, $dep(\mathbf{C}_j) = \mathbf{C}_j \cup (\mathbf{Pa}(\mathbf{C}_j) \setminus \mathbf{T})$ and $Y_j(\cdot) = 1$ (a constant function). Set $\mathbf{b} = (b_1, \ldots, b_p)$, $\mathbf{dep} := (dep(1), \ldots, dep(r), dep(\mathbf{C}_1), \ldots, dep(\mathbf{C}_p))$, and $\mathbf{Y} = (Y_1, \ldots, Y_p)$.

In Definition 6.1, we define an optimization program, $P_{r,p}(\Sigma, \gamma, \mathcal{C}, \mathbf{b}, \mathbf{dep}, \mathbf{Y})$ whose objective value is equal to $BC(P_{\mathcal{M}}[\mathbf{W} \mid do(\mathbf{t})], P_{\mathcal{N}}[\mathbf{W} \mid do(\mathbf{t})])$. In Section 6, we provide the steps to prove a lower bound on the objective of the program, thereby proving a lower bound on $BC(P_{\mathcal{M}}[\mathbf{W} \mid do(\mathbf{t})], P_{\mathcal{N}}[\mathbf{W} \mid do(\mathbf{t})])$.

Also, from (2) and (4), for all $j \in [p]$ and for all $\mathbf{w}_{dep(\mathbf{C}_j) \setminus \mathbf{C}_j}$,

$$\sum_{w_{n_{j,1}}} X_{n_{j,1}}(\mathbf{w}_{dep(n_{j,1})}) \sum_{w_{n_{j,2}}} X_{n_{j,2}}(\mathbf{w}_{dep(n_{j,2})}) \ldots \sum_{w_{n_{j,s_j}}} X_{n_{j,s_j}}(\mathbf{w}_{dep(n_{j,s_j})}) \geqslant 1 - \gamma$$

satisfying (6). Note that $P_{r,p}(\Sigma, \gamma, \mathcal{C}, \mathbf{b}, \mathbf{dep}, \mathbf{Y})$ is a program such that $\max_j |dep(C_j)| \leqslant \ell(d+1)$. By Lemma 6.6,

$$BC\left(P_{\mathcal{M}}[\mathbf{W} \mid do(\mathbf{t})], P_{\mathcal{N}}[\mathbf{W} \mid do(\mathbf{t})]\right) \geqslant \mathsf{Opt}(P_{r,p}) \geqslant (1 - |\Sigma|^{\ell(d+1)}\gamma)^p.$$

Using this in (4), we get

$$H^2\left(P_{\mathcal{M}}[\mathbf{W} \mid do(\mathbf{t})], P_{\mathcal{N}}[\mathbf{W} \mid do(\mathbf{t})]\right) \leqslant 1 - (1 - |\Sigma|^{\ell(d+1)}\gamma)^p \leqslant p\gamma|\Sigma|^{\ell(d+1)} \leqslant \varepsilon. \qquad \square$$

5 Lower Bound on Interventional Complexity

Recall that in Section 3 we provided non-adaptive algorithms for $\mathsf{C2ST}(G, \varepsilon)$, and $\mathsf{CL}(G, \varepsilon)$. In this section we provide lower bounds on the number of interventions that any algorithm must make to solve these problems. Our lower bounds nearly match the upper bounds in Theorem 3.2, and Theorem 3.4, even when the algorithm is allowed to be adaptive (namely future interventions are decided based upon the samples observed from the past interventions). In other words, these lower bounds show that adaptivity cannot reduce the interventional complexity.

Theorem 5.1. *There exists a* SMCG $G \in \mathcal{G}_{d,\ell}$ *with* n *nodes such that* $\Omega(K^{\ell d-2} \log n)$ *interventions are necessary for any algorithm (even adaptive) that solves* $\mathsf{C2ST}(G, \varepsilon)$ *or* $\mathsf{CL}(G, \varepsilon)$.

This theorem is proved via the following ingredients.

Necessary Condition. We obtain a necessary condition on the set of interventions $\mathbf{I}$ of any algorithm that solves $\mathsf{C2ST}(G, \varepsilon)$ or $\mathsf{CL}(G, \varepsilon)$.

We will consider SMCGs G with a specific structure, and prove the necessary condition for these graphs: The vertices of G are the union of two disjoint sets $\mathbf{A}$, and $\mathbf{B}$, such that G contains directed edges from $\mathbf{A}$ to $\mathbf{B}$, and bidirected edges within $\mathbf{B}$. Further, all edges in G are one of these two types. The next lemma, proved later in the section, is for graphs with this structure.

Lemma 5.2. *Suppose an adaptive algorithm uses a sequence of interventions $\mathbf{I}$ to solve $\mathsf{C2ST}(G, \varepsilon)$ or $\mathsf{CL}(G, \varepsilon)$. Let $\mathbf{C} \subseteq \mathbf{B}$ be a c-component of G. Then, for any assignment $\mathbf{pa}(\mathbf{C}) \in \Sigma^{|\mathbf{Pa}(\mathbf{C})|}$, there is an intervention $I \in \mathbf{I}$ such that the following conditions hold:*

C1. *I intervenes $\mathbf{Pa}(\mathbf{C})$ with the corresponding assignment of $\mathbf{pa}(\mathbf{C})$,*[5]

C2. *I does not intervene any node in $\mathbf{C}$.*

Existence. We then show that there is a graph with the structure mentioned above for which $\mathbf{I}$ must be $\Omega(K^{\ell d-2} \log n)$ in order for the condition to be satisfied. More precisely,

Lemma 5.3. *There exists a G, and a constant c such that for any set of interventions $\mathbf{I}$ with $|\mathbf{I}| < c \cdot K^{\ell d-2} \log n$, there is a $\mathbf{C} \subseteq \mathbf{B}$, which is a c-component of G, and an assignment $\mathbf{pa}(\mathbf{C})$ such that no intervention in $\mathbf{I}$*

- *assigns $\mathbf{pa}(\mathbf{C})$ to $\mathbf{Pa}(\mathbf{C})$, and*
- *observes all variables in $\mathbf{C}$.*

Proof of Lemma 5.3. We show existence of such a G using a probabilistic argument. We consider $\mathbf{A} = \mathbf{A}_r \cup \mathbf{A}_f$, where $\mathbf{A}_r := \{A_1, \ldots, A_n\}$, and $\mathbf{A}_f := \{A_{n+1}, \ldots, A_{n+(\ell d)-2}\}$. We consider $\mathbf{B} := \mathbf{B}_1 \cup \mathbf{B}_2 \cup \ldots \cup \mathbf{B}_{n/\ell}$, where for each $i \in [n/\ell]$, $\mathbf{B}_i = \{B_{i,1}, B_{i,2}, \ldots, B_{i,\ell}\}$. $\mathbf{V} = \mathbf{A} \cup \mathbf{B}$ will be the set of observable nodes in the graph. Therefore, the number of nodes is $|\mathbf{V}| = 2n + \ell d - 2 = O(n)$.

The set of unobservable nodes are such that the following is satisfied:

- $\mathbf{B}_i$ is a c-component in G, for each $\mathbf{B}_i$.

We consider random directed bipartite graphs on $\mathbf{V}$ generated as follows, where all the edges go from $\mathbf{A}$ to $\mathbf{B}$. Each c-component $\mathbf{B}_i$ has exactly ℓd parents, chosen as follows:

- $\mathbf{A}_f \subset \mathbf{Pa}(\mathbf{B}_i)$, namely every vertex of $\mathbf{A}_f$ is the parent of at least one node in $\mathbf{B}_i$.
- The remaining two parents of $\mathbf{B}_i$ are chosen randomly from $\mathbf{A}_r$ with edge density $p := 2/n$.

Let $\mathbf{I}$ be a set of interventions that satisfies the conditions of Lemma 5.2. Let $\mathbf{I}' \subseteq \mathbf{I}$ be the interventions that intervene *all the nodes* in $\mathbf{A}$. The nodes in $\mathbf{A}_f$ can be intervened in $|\Sigma|^{|\mathbf{A}_f|} = K^{\ell d-2}$ ways. This induces a partition of $\mathbf{I}'$ into $K^{\ell d-2}$ parts, where the interventions in each partition intervenes $\mathbf{A}_f$ with the same assignment. Let $\{\mathbf{I}_1, \ldots, \mathbf{I}_{K^{\ell d-2}}\}$ such that $\mathbf{I}' = \mathbf{I}_1 \cup \ldots \cup \mathbf{I}_{K^{\ell d-2}}$ be this partition. We will show that for each j, $|\mathbf{I}_j| = \Omega(\log n)$, implying that

$$|\mathbf{I}| \geqslant |\mathbf{I}'| \geqslant K^{\ell d-2} \cdot \Omega(\log n) = \Omega(K^{\ell d-2} \log n).$$

[5] In our construction, $\mathbf{Pa}(C)$ always take $\mathbf{0}$ in the natural distribution. Henceforth, the interventions where some vertices in $\mathbf{Pa}(C)$ are not intervened are not considered here, as they are equivalent to the case when those vertices are intervened with $\mathbf{0}$.

Consider a $\mathbf{I}_j$, with $|\mathbf{I}_j| = t$. Further, for simplicity we assume that $K = 2$ for this part, and that $\Sigma = \{0, 1\}$. Since all the nodes in $\mathbf{A}_r$ are intervened, consider one such node. For any node in $\mathbf{A}_r$ consider the t bit binary string denoting whether it is intervened with 0 or 1 in the t interventions. This divides the set $\mathbf{A}_r$ into 2^t cells $\mathbf{Z}_1, \ldots, \mathbf{Z}_{2^t}$, where two nodes are in the same cell if they are intervened with the identical value by each intervention in $\mathbf{I}_j$. The expected number of pairs of vertices in $\mathbf{Z}_h$ that are both parents of some vertex in $\mathbf{B}$ is $O(p|\mathbf{Z}_h|^2)$. Therefore, the expected number of pairs of vertices that are both parents of some vertex in $\mathbf{B}$ and also belong to the same cell is $O\left(\sum_h p|\mathbf{Z}_h|^2\right)$, which is at least $O(pn^2 2^{-t})$ (since $\sum_h \mathbf{Z}_h = n$). Now for any such pair of vertices $A, A' \in \mathbf{A}_r$ that belong to the same cell, there exists no intervention such that $A = 0$ and $A' = 1$, contradicting to our requirement. Therefore, $pn^2 2^{-t} < 1$ which implies t is at least $\Omega(\log n)$. □

Combining these two lemmas, we obtain the lower bound for the adaptive versions of $\mathsf{C2ST}(G, \varepsilon)$ and $\mathsf{CL}(G, \varepsilon)$. Now we proceed to prove Lemma 5.2.

Proof of Lemma 5.2. In our construction we consider models where $\mathbf{A}$ is assigned $\mathbf{0}^{|\mathbf{A}|}$ with probability one in the observable distribution. In other words, each $A_i \in \mathbf{A}$ takes value 0 with probability one. Consider any intervention I that targets a $\mathbf{A}' \subseteq \mathbf{A}$. Consider the intervention I' that intervenes $\mathbf{A}'$ the same way as I, but intervenes the nodes in $\mathbf{A} \setminus \mathbf{A}'$ with 0's. Since there are no incoming arrows to $\mathbf{A}$, the distribution of $\mathbf{I}'$ will be the same as $\mathbf{I}$. Therefore, we assume that each intervention I we make intervenes *all the vertices* in $\mathbf{A}$.

Suppose there is an algorithm that makes a series of interventions $\mathbf{I}$ that do not satisfy the conditions of Lemma 5.2. In other words, there exists a c-component $\mathbf{C} \subseteq \mathbf{B}$ and an assignment $\mathbf{pa}(\mathbf{C})$, such that no intervention in $\mathbf{I}$ satisfies **C1** and **C2**. Let $\mathbf{C} = \{V_1, V_2, \ldots, V_\ell\}$ and $\mathbf{Pa}(\mathbf{C}) = \{W_1, W_2, \ldots, W_s\}$.

Let G' be a subgraph of G on the vertices $\mathbf{C} \cup \mathbf{Pa}(\mathbf{C})$ whose edge set satisfies the following:

- $\mathbf{C}$ contains exactly $\ell - 1$ bidirected edges that form a tree.
- each of the parent vertices W_i has exactly one child node in $\mathbf{C}$.

In our construction, we consider models where the distribution on the rest of the vertices of G (*i.e.,* $\mathbf{V} \setminus (\mathbf{C} \cup \mathbf{Pa}(\mathbf{C}))$) will be independent of the distribution on $\mathbf{C} \cup \mathbf{Pa}(\mathbf{C})$. Therefore, we can restrict our focus on G'. We will show the existence of two models $\mathcal{M}$ and $\mathcal{N}$ on G' such that:

S.1 Let $\mathbf{T} \subseteq (\mathbf{C} \cup \mathbf{Pa}(\mathbf{C}))$, $\mathbf{t} \in \Sigma^{|\mathbf{T}|}$. Let $\{\mathbf{C}_1, \ldots, \mathbf{C}_q\}$ be the c-components of the induced graph $G'[\mathbf{C} \setminus \mathbf{T}]$. Suppose under the intervention $do(\mathbf{t})$, the conditions **C1**, or **C2** is not satisfied, then, the distributions over $\mathbf{C} \setminus \mathbf{T}$ in $\mathcal{M}$ and $\mathcal{N}$ are identical under $do(\mathbf{t})$, namely,

$$P_{\mathcal{M}}[\mathbf{C} \setminus \mathbf{T} \mid do(\mathbf{t})] = \prod_i P_{\mathcal{M}}[\mathbf{C}_i \mid do(\mathbf{t})] = \prod_i P_{\mathcal{N}}[\mathbf{C}_i \mid do(\mathbf{t})] = P_{\mathcal{N}}[\mathbf{C} \setminus \mathbf{T} \mid do(\mathbf{t})]$$

where for each i, $P_{\mathcal{M}}[\mathbf{C}_i \mid do(\mathbf{t})] = P_{\mathcal{N}}[\mathbf{C}_i \mid do(\mathbf{t})]$ and is a uniform distribution over $\{0, 1\}^{|\mathbf{C}_i|}$,

S.2 $\delta_{TV}(P_{\mathcal{M}}[\mathbf{C} \mid do(\mathbf{pa}(\mathbf{C}))], P_{\mathcal{N}}[\mathbf{C} \mid do(\mathbf{pa}(\mathbf{C}))]) = 1$.[6]

Recall that the sequence of interventions performed by an (adaptive) algorithm is denoted by $\mathbf{I}$. The assignment $\mathbf{pa}(\mathbf{C})$ gets fixed only after the algorithm fixes *all* the interventions in $\mathbf{I}$. However, we know

[6]Recall that $\mathbf{pa}(\mathbf{C})$ is the assignment that gets fixed after the algorithm fixes the sequence $\mathbf{I}$.

that any intervention in $\mathbf{I}$ belongs to the category S.1. And for each such intervention in $\mathbf{I}$, the corresponding distributions on models $\mathcal{M}$ and $\mathcal{N}$ are equal, and is defined by a set of uniform distributions over the c-components. Therefore, we can construct an adversary that, for each intervention in $\mathbf{I}$ performed by the algorithm (sequentially), outputs a distribution[7] based on S.1. When the algorithm terminates, the assignment $\mathbf{pa}(\mathbf{C})$ gets fixed, and we can show the existence of two models $\mathcal{M}$ and $\mathcal{N}$ such that

- the models agree on all the interventional distributions in $\mathbf{I}$, and all such distributions also match the corresponding distributions that were revealed by the adversary.
- $\delta_{TV}(P_{\mathcal{M}}[\mathbf{C} \mid do(\mathbf{pa}(\mathbf{C}))], P_{\mathcal{N}}[\mathbf{C} \mid do(\mathbf{pa}(\mathbf{C}))]) = 1$.

Moreover, we can construct such an adversary that outputs distributions in the same way, for all the c-components $\mathbf{C} \subseteq \mathbf{B}$ of G. Thus, an explicit construction of two models $\mathcal{M}$ and $\mathcal{N}$ on G' that generates distributions according to S.1 and S.2 would conclude our proof. The remainder of the proof is dedicated towards this goal.

Let $\mathbf{U}$ be the set of all unobservable variables in G'. Let $\mathbf{U}^{V_i} \subseteq \mathbf{U}$ represent the bidirected edges incident to V_i in G'. Also, for each variable V_i we have an additional boolean random variable R_i that provides randomness to V_i. All the randomness in the models $\mathcal{M}$ and $\mathcal{N}$ we construct are in the hidden variables U_i's and the R_i's. In other words, the observable variables are a deterministic function of these. The models $\mathcal{M}$ and $\mathcal{N}$ are defined as follows:

1. (a) For each bidirected edge $U_i \in \mathbf{U}$, U_i is a Bern(0.5) random variable in both $\mathcal{M}$, and $\mathcal{N}$.
   (b) In each model, R_i's are also independent Bern(0.5) random variables.
2. For each $i \in [s]$, $W_i = 0$ with probability one in both $\mathcal{M}$, and $\mathcal{N}$.
3. For each $V_i \in \mathbf{C}$, with probability one:
   (a) when $\mathbf{Pa}(V_i)$ is not consistent with $\mathbf{pa}(V_i)$, then $V_i = \mathsf{XOR}(\mathbf{U}^{V_i}, R_i)$ in both both $\mathcal{M}$, and $\mathcal{N}$.
   (b) when $\mathbf{Pa}(V_i)$ is consistent with $\mathbf{pa}(V_i)$ and $i \neq 1$, then $V_i = \mathsf{XOR}(\mathbf{U}^{V_i})$ in both $\mathcal{M}$, and $\mathcal{N}$.
   (c) when $\mathbf{Pa}(V_i)$ is consistent with $\mathbf{pa}(V_i)$ and $i = 1$, V_i takes
   - $V_i = \mathsf{XOR}(\mathbf{U}^{V_i})$ in $\mathcal{M}$, and
   - $V_i = \mathsf{XNOR}(\mathbf{U}^{V_i})$ in $\mathcal{N}$.

Case 1: When I respects S.1. Consider an intervention I, identified by $do(\mathbf{t})$, that respects S.1. That is, either I intervenes some node in $\mathbf{C}$, or I does not intervene $\mathbf{Pa}(\mathbf{C})$ with the assignment $\mathbf{pa}(\mathbf{C})$. Let $\{\mathbf{C}_1, \ldots, \mathbf{C}_q\}$ be the c-components of the graph induced by $\mathbf{C} \setminus \mathbf{T}$. Note that the models M, and N, differ only on the function V_1. Therefore, when V_1 is intervened in I, it is easy to see that the required distributions are equal, and is a product of uniform distributions over the c-components.[8] Suppose V_1 is not intervened in I, and without loss of generality let $\mathbf{C}_1$ be the c-component that contains V_1. Since the models differ only on V_1, it is easy to see that $P_{\mathcal{M}}[\mathbf{C}_i \mid do(\mathbf{t})] = P_{\mathcal{N}}[\mathbf{C}_i \mid do(\mathbf{t})]$ for all $i \neq 1$, and is uniform over $\{0,1\}^{|\mathbf{C}_i|}$. Hence, it is sufficient to prove that $P[\mathbf{C}_1 \mid do(\mathbf{t})]$'s are equal and uniform in both models. Let $\mathbf{S}$ be the set of all U_i's and R_j's of the following type: a) U_i's that have one child in $\mathbf{C}_1$ and another child in $\mathbf{T}$; b) R_j's with

[7] We consider the worst case, where the algorithm is provided with infinite samples.

[8] Recall that our objective is to prove: $P_{\mathcal{M}}[\mathbf{C} \setminus \mathbf{T} \mid do(\mathbf{t})] = \prod_i P_{\mathcal{M}}[\mathbf{C}_i \mid do(\mathbf{t})] = \prod_i P_{\mathcal{N}}[\mathbf{C}_i \mid do(\mathbf{t})] = P_{\mathcal{N}}[\mathbf{C} \setminus \mathbf{T} \mid do(\mathbf{t})]$, where for each i, $P_{\mathcal{M}}[\mathbf{C}_i \mid do(\mathbf{t})] = P_{\mathcal{N}}[\mathbf{C}_i \mid do(\mathbf{t})]$ is a uniform distribution over $\{0,1\}^{|\mathbf{C}_i|}$.

respect to $V_j \in \mathbf{C}_1$ such that $\mathbf{pa}(V_j)$[9] is inconsistent with $\mathbf{t}$ (*i.e.,* $V_j \in \mathbf{C}_1$ that computes $\mathsf{XOR}(\mathbf{U}^{V_j}, R_j)$). Now, for any fixed assignment $\mathbf{a}_{\mathbf{C}_1^{-1}}$ to $\mathbf{C}_1 \setminus \{V_1\}$ and $\mathbf{a}_\mathbf{S}$ to $\mathbf{S}$, because the bi-directed edges within $\mathbf{C}_1$ form a '*tree*', the value of every unobservable variable within $\mathbf{C}_1$[10] can be computed. Note that V_1 computes $\mathsf{XOR}(\mathbf{a}_{\mathbf{C}_1^{-1}}, \mathbf{a}_S)$ in $\mathcal{M}$, and $\mathsf{XNOR}(\mathbf{a}_{\mathbf{C}_1^{-1}}, \mathbf{a}_S)$ in $\mathcal{N}$. However, we know that $\mathbf{S}$ is a non-empty set and the bit parities of $\mathbf{S}$ are uniformly distributed in both the models. This implies $P_\mathcal{M}[\mathbf{C}_1 \mid do(\mathbf{t})] = P_\mathcal{N}[\mathbf{C}_1 \mid do(\mathbf{t})]$, and is a uniform distribution over $\{0,1\}^{|\mathbf{C}_1|}$.

Case 2: When I respects S.2. Consider an intervention I that respects S.2. That is, $\mathbf{Pa}(\mathbf{C})$ is intervened with the assignment $\mathbf{pa}(C)$ in I, and no node of $\mathbf{C}$ is intervened in I. Consider the set of variables $\mathbf{S}$ as defined before for the S.1 case. Note that $\mathbf{S}$ is empty here. This implies, for any fixed assignment $\mathbf{a}_{\mathbf{C}^{-1}}$ to $\mathbf{C} \setminus \{V_1\}$, V_1 computes $\mathsf{XOR}(\mathbf{a}_{\mathbf{C}^{-1}})$ in $\mathcal{M}$, and V_1 computes $\mathsf{XNOR}(\mathbf{a}_{\mathbf{C}^{-1}})$ in $\mathcal{N}$. This implies, the supports of $P_\mathcal{M}[\mathbf{C} \mid do(\mathbf{pa}(\mathbf{C}))]$ and $P_\mathcal{N}[\mathbf{C} \mid do(\mathbf{pa}(\mathbf{C}))]$ are disjoint, and therefore the total variation distance is 1.

Hence, irrespective of the number of samples taken from the interventions of $\mathbf{I}$, any adaptive algorithm that solves $\mathsf{C2ST}(G, \varepsilon)$ or $\mathsf{CL}(G, \varepsilon)$ must consider a sequence of interventions that satisfies the conditions **C1** and **C2**.

□

6 Program $P_{r,p}$ and Properties

In this section, we gather the technical tools used to prove the subadditivity result, Theorem 4.5. We formulate our claims at a higher level of abstraction than needed for our purposes, so that the essence of the argument becomes clearer.

We begin by defining the optimization problem, and then describe it at a high level.

Definition 6.1 (Program $P_{r,p}(\Sigma, \gamma, \mathcal{C}, \mathbf{b}, \mathbf{dep}, \mathbf{Y})$). *For integers $r, p \geqslant 0$, suppose the following are given:*

1. *an alphabet set Σ,*
2. *$\gamma \in (0,1)$,*
3. *a partition[11] $\mathcal{C}$ of $[r]$ into $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_p$, where for each $j \in [p]$, $s_j = |\mathbf{C}_j|$ and the elements of $\mathbf{C}_j$ are $\{n_{j,1}, \ldots, n_{j,s_j}\}$ in increasing order,*
4. *a vector $\mathbf{b} = (b_1, b_2, \ldots, b_p) \in [0,1]^p$,*
5. *a vector of sets $\mathbf{dep} = (dep(1), \ldots, dep(r), dep(\mathbf{C}_1), \ldots, dep(\mathbf{C}_p))$ such that:*

$$\begin{aligned} [n_{j,i}] \supseteq dep(n_{j,i}) \supseteq \{n_{j,i}\} \cup dep(n_{j,i-1}) \qquad & \forall j \in [p], i \in [s_j] \\ s_j \neq 0 \implies dep(\mathbf{C}_j) \supseteq dep(n_{j,s_j}) \qquad & \forall j \in [p] \\ s_j = 0 \implies dep(\mathbf{C}_j) = \emptyset \qquad & \forall j \in [p] \end{aligned}$$

6. *a set of functions $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$, where $Y_j \colon \Sigma^{|dep(\mathbf{C}_j)|} \to [0,1]$.*

[9] We refer $\mathbf{pa}(V_j)$ with respect to the assignment $\mathbf{pa}(\mathbf{C})$.

[10] We refer to the unobservable variables U_i's where both the children of U_i lie in $\mathbf{C}_1$.

[11] Here, we allow some members of $\mathcal{C}$ to be empty sets.

The program $P_{r,p}(\Sigma, \gamma, \mathcal{C}, \mathbf{b}, \mathbf{dep}, \mathbf{Y})$ *is the following optimization problem over* $\mathbf{X} = (X_1, \ldots, X_r)$ *where* $X_i \colon \Sigma^{|dep(i)|} \to [0,1]$*:*

$$\min_{\mathbf{X}} f_{r,p}(\mathbf{X}) \stackrel{\text{def}}{=} \sum_{a_1 \in \Sigma} X_1(\mathbf{a}_{dep(1)}) \sum_{a_2 \in \Sigma} X_2(\mathbf{a}_{dep(2)}) \quad \cdots \sum_{a_r \in \Sigma} X_r(\mathbf{a}_{dep(r)}) \cdot \prod_{j=1}^{p} Y_j(\mathbf{a}_{dep(\mathbf{C}_j)})$$

subject to

$$\sum_{a_i \in \Sigma} X_i(\mathbf{a}_{dep(i)}) \leqslant 1 \qquad \forall i \in [r], \forall \mathbf{a}_{dep(i)\setminus\{i\}} \in \Sigma^{|dep(i)\setminus\{i\}|} \tag{5}$$

$$\sum_{a_{n_{j,1}} \in \Sigma} X_{n_{j,1}}(\mathbf{a}_{dep(n_{j,1})}) \sum_{a_{n_{j,2}} \in \Sigma} X_{n_{j,2}}(\mathbf{a}_{dep(n_{j,2})}) \cdots \sum_{a_{n_{j,s_j}} \in \Sigma} X_{n_{j,s_j}}(\mathbf{a}_{dep(n_{j,s_j})}) \cdot Y_j(\mathbf{a}_{dep(\mathbf{C}_j)})$$

$$\geqslant 1 - b_j\gamma \qquad \forall j \in [p], \forall \mathbf{a}_{dep(\mathbf{C}_j)\setminus\mathbf{C}_j} \tag{6}$$

The variables of this program are functions that are based on the Bhattacharya coefficients between distributions[12] on certain variables, and were described in Section 4.2. (5) captures the fact that the Bhattacharyya coefficient is at most one. (6) captures the closeness constraint in Theorem 4.5, *i.e.,* (2). Proving a lower bound on the objective value of this program will suffice to prove Theorem 4.5. The remainder of this section is dedicated towards this goal. Let $\mathsf{Opt}(P_{r,p})$ denote the optimal value of the program. The next three lemmas (Lemmas 6.2, 6.3 and 6.4) all have the following flavor:

- They take as input an optimization problem (program $P_{r,p}$), and output a new program $P^{\text{new}}_{r',p}$.
- The optimal value of the program only goes down.
- The new program is *simpler* to analyze.[13]

We pass the original program $P_{r,p}$ through the first lemma, and pass its output through the second. The second lemma is applied multiple times until the output program satisfies a particular property. The obtained program is then passed through the third lemma to obtain a new program $P_{r-1,p}$ (with a reduced value of r), and the steps repeat. The above procedure reduces to a program with $r = 0$, namely to a program of the form $P_{0,p}$. We can lower bound the objective of this program by simply using (6). Combining these will yield a lower bound on the optimum of the original program $P_{r,p}$, thus proving Theorem 4.5.

The first lemma takes a program as input and outputs a new program with a smaller optimal value that satisfies $dep(r) = dep(\mathbf{C}_f)$ (where $r \in \mathbf{C}_f$).

Lemma 6.2 (Dependent Set Reduction)**.** *Suppose* $r \in \mathbf{C}_f$*. Let* $P^{new}_{r,p}$ *be the program obtained from* $P_{r,p}$ *by replacing* $dep(r)$ *by* $dep(\mathbf{C}_f)$*, then*

$$\mathsf{Opt}(P_{r,p}) \geqslant \mathsf{Opt}(P^{new}_{r,p}).$$

Proof. Our goal is to reduce the given program $P_{r,p}$ to a different program $P^{\text{new}}_{r,p}$ such that $\mathsf{Opt}(P_{r,p}) \geqslant \mathsf{Opt}(P^{\text{new}}_{r,p})$, where $P^{\text{new}}_{r,p}$ is defined from $P_{r,p}$ by defining $dep(r)$ to be $dep(\mathbf{C}_f)$.

Let $\mathbf{X}^{\text{old}} = \{X^{\text{old}}_1, \ldots X^{\text{old}}_r\}$ be an optimal solution of $P_{r,p}$. Now we construct a *feasible* solution $\mathbf{X}^{\text{new}} = \{X^{\text{new}}_1, \ldots, X^{\text{new}}_r\}$ for the program $P^{\text{new}}_{r,p}$, such that $f^{\text{new}}_{r,p}(\mathbf{X}^{\text{new}}) = f_{r,p}(\mathbf{X}^{\text{old}}) = \mathsf{Opt}(P_{r,p})$. For all

[12] The distributions may be interventional, or conditional, or a combination of condional and interventional distributions.

[13] We understand that this item is very subjective.

$i \neq r$, we define $X_i^{\text{new}} = X_i^{\text{old}}$. For $i = r$, we define $X_r^{\text{new}}(\mathbf{a}_{dep^{\text{new}}(r)}) = X_r^{\text{old}}(x_{dep(r)})$. In other words, X_r^{new} ignores the new variables added to $dep^{\text{new}}(r)$. Therefore,

$$\begin{aligned} f_{r,p}^{\text{new}}(\mathbf{X}^{\text{new}}) &= \sum_{a_1 \in \Sigma} X_1^{\text{new}}(\mathbf{a}_{dep(1)}) \cdots \sum_{a_r \in \Sigma} X_r^{\text{new}}(\mathbf{a}_{dep^{\text{new}}(r)}) \cdot \prod_{j=1}^{p} Y_j(\mathbf{a}_{dep(\mathbf{C}_j)}) \\ &= \sum_{a_1 \in \Sigma} X_1^{\text{old}}(\mathbf{a}_{dep(1)}) \cdots \sum_{a_r \in \Sigma} X_r^{\text{old}}(\mathbf{a}_{dep(r)}) \cdot \prod_{j=1}^{p} Y_j(\mathbf{a}_{dep(\mathbf{C}_j)}) \quad \text{(by the definition of } \mathbf{X}^{\text{new}}) \\ &= f_{r,p}(X^{\text{old}}) \qquad \text{(by the definition of } f_{r,p}). \end{aligned}$$

For the program $P_{r,p}^{\text{new}}$, when $i \neq r$, $\mathbf{X}^{\text{new}}$ satisfies the constraints in (5) (since the functions X_i^{old} and X_i^{new} are the same). Similarly, for $j \neq f$, constraints in (6) of the program $P_{r,p}^{\text{new}}$ are valid. When $i = r$ in (5), for each $\mathbf{a}_{dep^{\text{new}}(r)\setminus\{r\}}$, we get

$$\sum_{a_r} X_r^{\text{new}}(\mathbf{a}_{dep^{\text{new}}(r)}) = \sum_{\mathbf{a}_r} X_r^{\text{old}}(\mathbf{a}_{dep(r)}) \leqslant 1.$$

When $j = f$ in (6), for all $\mathbf{a}_{dep(\mathbf{C}_j)\setminus\mathbf{C}_j}$, since $n_{f,s_f} = r$ we get,

$$\begin{aligned} &\sum_{a_{n_{f,1}} \in \Sigma} X_{n_{f,1}}^{\text{new}}(\mathbf{a}_{dep(n_{f,1})}) \cdots \sum_{a_r \in \Sigma} X_r^{\text{new}}(\mathbf{a}_{dep^{\text{new}}(r)}) \cdot Y_f(\mathbf{a}_{dep(\mathbf{C}_f)}) \\ &= \sum_{a_{n_{f,1}} \in \Sigma} X_{n_{f,1}}^{\text{old}}(\mathbf{a}_{dep(n_{f,1})}) \cdots \sum_{a_r \in \Sigma} X_r^{\text{old}}(\mathbf{a}_{dep(r)}) \cdot Y_f(\mathbf{a}_{dep(\mathbf{C}_f)}) \quad \text{(from definition of } \mathbf{X}^{\text{new}}) \\ &\geqslant 1 - b_f\gamma \qquad \text{(using (6)).} \end{aligned}$$

This implies X^{new} is a feasible solution for $P_{r,p}^{\text{new}}$ and hence $\mathsf{Opt}(P_{r,p}) \geqslant \mathsf{Opt}(P_{r,p}^{\text{new}})$. □

The next lemma takes a program $P_{r,p}$ as input and outputs a new program (with a smaller optimal value) that satisfies $r \notin dep(\mathbf{C}_h)$ (for some given $\mathbf{C}_h$ such that $r \notin \mathbf{C}_h$).

Lemma 6.3 (Y-R Reduction)**.** *Let $P_{r,p}(\Sigma, \gamma, \mathcal{C}, \mathbf{b}, \mathbf{dep}, \mathbf{Y})$ be a given program, and there exists $h \in [p]$ such that $r \notin \mathbf{C}_h$ and $r \in dep(\mathbf{C}_h)$. Then, there exists a program $P_{r,p}^{new}(\Sigma, \gamma, \mathcal{C}, \mathbf{b}^{new}, \mathbf{dep}^{new}, \mathbf{Y}^{new})$ such that*

$$\mathsf{Opt}(P_{r,p}) \geqslant \mathsf{Opt}(P_{r,p}^{new}),$$

where

1. $b_h^{new} = |\Sigma| \cdot b_h$
2. $dep^{new}(\mathbf{C}_h) = dep(\mathbf{C}_h) \setminus \{r\}$
3. $b_j^{new} = b_j \qquad \forall j \in [p] \setminus \{h\}$
4. $dep^{new}(C_j) = dep(C_j) \qquad \forall j \in [p] \setminus \{h\}$
5. $dep^{new}(i) = dep(i) \qquad \forall i \in [r]$
6. $Y_j^{new}(\mathbf{a}_{dep(\mathbf{C}_j)}) = Y_j(\mathbf{a}_{dep(\mathbf{C}_j)}) \quad \forall j \in [p] \setminus \{h\}, \forall \mathbf{a}_{dep(\mathbf{C}_j)}.$

Proof. Let $\mathbf{X}'$ be an optimal solution of $P_{r,p}$. Note that, since $dep(\mathbf{C}_h^{\text{new}}) = dep(\mathbf{C}_h) \setminus \{r\}$, our goal is to find a function $Y_h^{\text{new}} : \Sigma^{|dep(\mathbf{C}_h^{\text{new}})|} \to [0, 1]$, whose domain size is smaller than the domain size of Y_h (as Y_h^{new} is independent of the value of a_r), that satisfies the required constraints.

For a given set of functions $\mathbf{X}$, a subset $\mathbf{S} \subseteq [r]$, and for a given assignment $a_{\mathbf{S}}$ to $\mathbf{S}$, let $f_{r,p}(\mathbf{X})|_{a_{\mathbf{S}}}$ represent the sum of all terms in $f_{r,p}(\mathbf{X})$ that are consistent with the assignment $a_{\mathbf{S}}$. Note that

$$f_{r,p}(\mathbf{X}') = \sum_{\mathbf{a}_{dep(\mathbf{C}_h)}} f_{r,p}(\mathbf{X}')|_{\mathbf{a}_{dep(\mathbf{C}_h)}}. \tag{7}$$

For each $\mathbf{a}_{dep(\mathbf{C}_h)\setminus\{r\}}$, let

1. $z_h(\mathbf{a}_{dep(\mathbf{C}_h)\setminus\{r\}}) = \arg\min_{a_r} Y_h(\mathbf{a}_{dep(\mathbf{C}_h)\setminus\{r\}}, a_r)$,
2. $Y_h^{\text{new}}(\mathbf{a}_{dep^{\text{new}}(\mathbf{C}_h)}) = Y_h^{\text{new}}(\mathbf{a}_{dep(\mathbf{C}_h)\setminus\{r\}}) = Y_h(\mathbf{a}_{dep(\mathbf{C}_h)\setminus\{r\}}, z_h(\mathbf{a}_{dep(\mathbf{C}_h)\setminus\{r\}}))$.

Based on the above definition of Y_h^{new}, we know that $f_r^{\text{new}}(\mathbf{X}') \leqslant f_{r,p}(\mathbf{X}')$. In the remainder of the proof, we show that $\mathbf{X}'$ is also a feasible solution for $P_{r,p}^{\text{new}}$. The first set of constraints of $P_{r,p}^{\text{new}}$ are valid (as we have not modified $\mathbf{X}$). Similarly, the second set of constraints is valid for all $j \neq h$ (as we have not changed any parameters). Now we prove the constraints in (6), for $j = h$. For all assignments $a_{dep^{\text{new}}(\mathbf{C}_h)\setminus\mathbf{C}_h}$,

$$\begin{aligned}
&\sum_{a_{n_{h,1}}} X'_{n_{h,1}}(\mathbf{a}_{dep(n_{h,1})}) \cdots \sum_{a_{n_{h,s_h}}} X'_{n_{h,s_h}}(\mathbf{a}_{dep(n_{h,s_h})}) \cdot Y_h^{\text{new}}(\mathbf{a}_{dep^{\text{new}}(C_h)}) \\
&\quad = \sum_{a_{n_{h,1}}} X'_{n_{h,1}}(\mathbf{a}_{dep(n_{h,1})}) \cdots \sum_{a_{h_{s_j}}} X'_{n_{h,s_h}}(\mathbf{a}_{dep(n_{h,s_h})}) \cdot Y_h(\mathbf{a}_{dep(\mathbf{C}_h)\setminus\{r\}}, a_r = z_h(\mathbf{a}_{dep(\mathbf{C}_h)\setminus\{r\}})) \\
&\qquad\qquad \text{(by definition of } Y_h^{\text{new}}) \\
&\quad = \left[\sum_{a_{n_{h,1}}} X'_{n_{h,1}}(\mathbf{a}_{dep(n_{h,1})}) \cdots \sum_{a_{h_{s_j}}} X'_{n_{h,s_h}}(\mathbf{a}_{dep(n_{h,s_h})}) \cdot \sum_{a_r} Y_h(\mathbf{a}_{dep(\mathbf{C}_h\setminus\{r\})}, a_r)\right] \\
&\qquad - \left[\sum_{a_{n_{h,1}}} X'_{n_{h,1}}(\mathbf{a}_{dep(n_{h,1})}) \cdots \sum_{a_{h_{s_j}}} X'_{n_{h,s_h}}(\mathbf{a}_{dep(n_{h,s_h})}) \cdot \sum_{a_r : a_r \neq z_h(\mathbf{a}_{dep(\mathbf{C}_h)\setminus\{r\}})} Y_h(\mathbf{a}_{dep(\mathbf{C}_h)\setminus\{r\}}, a_r)\right] \\
&\quad \geqslant \left[\sum_{a_r}\sum_{a_{n_{h,1}}} X'_{n_{h,1}}(\mathbf{a}_{dep(n_{h,1})}) \cdots \sum_{a_{n_{h,s_h}}} X'_{n_{h,s_h}}(\mathbf{a}_{dep(n_{h,s_h})}) \cdot Y_h(\mathbf{a}_{dep(\mathbf{C}_h)\setminus\{r\}}, a_r)\right] \\
&\qquad - \left[\sum_{a_{n_{h,1}}} X'_{n_{h,1}}(\mathbf{a}_{dep(n_{h,1})}) \cdots \sum_{a_{h_{s_j}}} X'_{n_{h,s_h}}(\mathbf{a}_{dep(n_{h,s_h})}) \cdot \sum_{a_r : a_r \neq z(\mathbf{a}_{dep(\mathbf{C}_h)\setminus\{r\}})} 1\right] \quad (\because Y_h(.) \leqslant 1) \\
&\quad \geqslant \left[\sum_{a_r \in \Sigma} (1 - b_h\gamma)\right] - \left[(|\Sigma - 1|) \cdot \sum_{a_{n_{h,1}}} X'_{n_{h,1}}(\mathbf{a}_{dep(n_{h,1})}) \cdots \sum_{a_{h_{s_j}}} X'_{n_{h,s_h}}(\mathbf{a}_{dep(n_{h,s_h})})\right] \\
&\qquad\qquad \text{(by constraint (6) of } P_{r,p}) \\
&\quad \geqslant |\Sigma|(1 - b_h\gamma) - (|\Sigma| - 1)1 \quad \text{(by constraint (5) of } P_{r,p}) \\
&\quad = 1 - |\Sigma| b_h \gamma \\
&\quad = 1 - b_h^{\text{new}}\gamma.
\end{aligned}$$

□

After multiple passes through the above lemma, we get a program $P_{r,p}$ that satisfies $r \notin dep(\mathbf{C}_j)$, for all $\mathbf{C}_j$ such that $r \notin \mathbf{C_j}$. The next lemma takes in such a program, and outputs a program with a reduced value of r.

Lemma 6.4 (R-Elimination)**.** *Let $P_{r,p}(\Sigma, \gamma, \mathcal{C}, \mathbf{b}, \mathbf{dep}, \mathbf{Y})$ be a given program such that the element $r \in \mathbf{C}_f$. Suppose $dep(r) = dep(\mathbf{C}_f)$, and for all $j \in [p] \setminus \{f\}$, $r \notin dep(\mathbf{C}_j)$. Then there exists a program $P^{new}_{r-1,p}(\Sigma, \gamma, \mathcal{C}^{new}, \mathbf{b}, \mathbf{dep}^{new}, \mathbf{Y}^{new})$ such that*

$$\mathsf{Opt}(P_{r,p}) \geqslant \mathsf{Opt}(P^{new}_{r-1,p})$$

where $\mathbf{Y}^{new}$ differs from $\mathbf{Y}$ only on the function Y_f, $\mathcal{C}^{new}$ differs from $\mathcal{C}$ only on the partition $\mathbf{C}_f$ where $\mathbf{C}^{new}_f = \mathbf{C}_f \setminus \{r\}$, and $\mathbf{dep}^{new} = (dep(1), dep(2), \ldots, dep(r-1), dep(\mathbf{C}_1), \ldots, dep(\mathbf{C}_{f-1}), dep^{new}(\mathbf{C}^{new}_f), dep(\mathbf{C}_{f+1}), \ldots, dep(\mathbf{C}_p))$ where $dep^{new}(\mathbf{C}^{new}_f) = dep(r) \setminus \{r\}$.

Proof. Let $\mathbf{X}^{\text{old}}$ be an optimal solution of $P_{r,p}$. For a given set of functions $\mathbf{X}$, a subset $\mathbf{S} \subseteq [r]$, and for a given assignment $a_{\mathbf{S}}$ to $\mathbf{S}$, let $f_{r,p}(\mathbf{X})|_{a_{\mathbf{S}}}$ represent the sum of all terms in $f_{r,p}(\mathbf{X})$ that are consistent with the assignment $a_{\mathbf{S}}$. Then, for all assignments $\mathbf{a}_{dep(r)\setminus\{r\}}$

$$f_{r,p}(\mathbf{X}^{\text{old}})|_{\mathbf{a}_{dep(r)\setminus\{r\}}} = L_{\mathbf{a}_{dep(r)\setminus\{r\}}} \cdot \sum_{a_r} X^{\text{old}}_r(\mathbf{a}_{dep(r)}) \cdot Y_f(\mathbf{a}_{dep(r)}) \tag{8}$$

We define $Y^{\text{new}}_f(\mathbf{a}_{dep^{\text{new}}(\mathbf{C}^{\text{new}}_f)}) = Y^{\text{new}}_f(\mathbf{a}_{dep(r)\setminus\{r\}}) = \sum_{a_r} X^{\text{old}}_r(\mathbf{a}_{dep(r)}) \cdot Y_f(\mathbf{a}_{dep(r)})$. Observe that $Y^{\text{new}}_f : \Sigma^{|dep(r)\setminus\{r\}|} \to [0,1]$ because of constraint (5) and since Y_f itself falls in the range $[0,1]$. Now, the new program $P^{\text{new}}_{r-1,p}$ is completely specified.

Observe that:

$$f_{r,p}(\mathbf{X}^{\text{old}})|_{\mathbf{a}_{dep(r)\setminus\{r\}}} = L_{\mathbf{a}_{dep(r)\setminus\{r\}}} \cdot Y^{\text{new}}_f(\mathbf{a}_{dep(r)\setminus\{r\}}) = f^{\text{new}}_{r-1,p}(\mathbf{X}^{\text{old}}_{r-1})|_{\mathbf{a}_{dep(r)\setminus\{r\}}}$$

where $\mathbf{X}^{\text{old}}_{r-1} = \{X^{\text{old}}_1, \ldots, X^{\text{old}}_{r-1}\}$. This implies

$$f_{r,p}(\mathbf{X}^{\text{old}}) = \sum_{\mathbf{a}_{dep(r)\setminus\{r\}}} f_{r,p}(\mathbf{X}^{\text{old}})|_{\mathbf{a}_{dep(r)\setminus\{r\}}} = \sum_{\mathbf{a}_{dep(r)\setminus\{r\}}} f^{\text{new}}_{r-1,p}(\mathbf{X}^{\text{old}}_{r-1})|_{\mathbf{a}_{dep(r)\setminus\{r\}}} = f^{\text{new}}_{r-1,p}(\mathbf{X}^{\text{old}}_{r-1}).$$

We now show that the functions $\mathbf{X}^{\text{old}}_{r-1}$ form a feasible solution for $P^{\text{new}}_{r-1,p}$. The first set of constraints (5) holds for $P^{\text{new}}_{r-1,p}$ because $\mathbf{X}^{\text{old}}$ is feasible for $P_{r,p}$. Also for all $j \neq f$, the second set of constraints (6) holds for the same reason. For $j = f$:

$$\begin{aligned}
&\sum_{a_{n_{f,1}} \in \Sigma} X^{\text{old}}_{n_{f,1}}(\mathbf{a}_{dep(n_{f,1})}) \cdots \sum_{a_{n_{f,s_f-1}} \in \Sigma} X^{\text{old}}_{n_{f,s_f-1}}(\mathbf{a}_{dep(n_{f,s_f-1})}) \cdot Y^{\text{new}}_f(\mathbf{a}_{dep(C^{\text{new}}_f)}) \\
&= \sum_{a_{n_{f,1}} \in \Sigma} X^{\text{old}}_{n_{f,1}}(\mathbf{a}_{dep(n_{f,1})}) \cdots \sum_{a_{n_{f,s_f-1}} \in \Sigma} X^{\text{old}}_{n_{f,s_f-1}}(\mathbf{a}_{dep(n_{f,s_f-1})}) \sum_{a_r \in \Sigma} X^{\text{old}}_r(\mathbf{a}_{dep(r)}) \cdot Y_f(\mathbf{a}_{dep(\mathbf{C}_f)}) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by definition)} \\
&\geqslant 1 - b_f\gamma
\end{aligned}$$

This completes the proof that $\mathsf{Opt}(P_{r,p}) \geqslant \mathsf{Opt}(P^{\text{new}}_{r-1,p})$. □

Lemma 6.5. *For any integers $r, p \geqslant 1$ and given a program $P_{r,p}(\Sigma, \gamma, \mathcal{C}, \mathbf{b}, \mathbf{dep}, \mathbf{Y})$, there exists a program $P^{new}_{r-1,p}(\Sigma, \gamma, \mathcal{C}^{new}, \mathbf{b}^{new}, \mathbf{dep}^{new}, \mathbf{Y}^{new})$ such that*

$$\mathsf{Opt}(P_{r,p}) \geqslant \mathsf{Opt}(P^{new}_{r-1,p})$$

where

$$\begin{aligned} b_j^{new} &= b_j && , \forall j \in [p] \colon r \notin dep(\mathbf{C}_j), \\ b_j^{new} &= |\Sigma| \cdot b_j && , \forall j \in [p] \colon r \in dep(\mathbf{C}_j). \end{aligned}$$

Proof. First we apply Lemma 6.2 (Dependent Set Reduction). Then, we apply Lemma 6.3 (Y-R Reduction) repeatedly, until there does not exist any $h \in [p]$ such that $r \in dep(\mathbf{C}_h)$ but $r \notin \mathbf{C}_h$. Note that, in each step of this reduction, the respective b_h increases by a factor of $|\Sigma|$. Finally, applying Lemma 6.4 (R-Elimination) results in a program $P^{\text{new}}_{r-1,p}$ on $r-1$ inputs with the desired property. □

Lemma 6.6. *For a given program $P_{r,p}(\Sigma, \gamma, \mathcal{C}, \mathbf{b} = \mathbf{1}, \mathbf{dep}, \mathbf{Y} = \mathbf{1})$, suppose we know that $\max_j |dep(\mathbf{C}_j)|$ is at most L. Then*

$$\mathsf{Opt}(P_{r,p}) \geqslant (1 - |\Sigma|^L \gamma)^p.$$

Proof. We apply Lemma 6.5 recursively. Note that in each such reduction from $P_{r,p}$ to $P_{r-1,p}$, the value of b_j increases by a factor of $|\Sigma|$ only when $r \in dep(\mathbf{C}_j)$.

At $r = 0$, we have the program $P_{0,p}(\Sigma, \gamma, \mathcal{C}', \mathbf{b}', \mathbf{dep}', \mathbf{Y}')$. For all $j \in [p]$, we know that $b'_j \leqslant |\Sigma|^L$ (since $|dep(\mathbf{C}_j)| \leqslant L$). Therefore,

$$\begin{aligned} \mathsf{Opt}(P_{r,p}) &\geqslant \mathsf{Opt}(P_{0,p}) \\ &= \prod_{j=1}^{p} Y'_j(\emptyset) \\ &\geqslant \prod_{j=1}^{p} (1 - b'_j \gamma) \quad \text{based on constraint (6) of the program } P_{0,p} \\ &\geqslant (1 - |\Sigma|^L \gamma)^p. \end{aligned}$$

□

7 Acknowledgments

We would like to thank Vasant Honavar who told us about the problems considered here and for several helpful discussions that were essential for us to complete this work.

References

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 3577–3598. Curran Associates, Inc., 2015. 6

[AGHP92] Noga Alon, Oded Goldreich, Johan Håstad, and René Peralta. Simple constructions of almost k-wise independent random variables. *Random Structures & Algorithms*, 3(3):289–304, 1992. 16

[AKN06] Pieter Abbeel, Daphne Koller, and Andrew Y. Ng. Learning factor graphs in polynomial time and sample complexity. *Journal of Machine Learning Research*, 7(Aug):1743–1788, 2006. 6

[ARSZ05] R Ayesha Ali, Thomas S Richardson, Peter Spirtes, and Jiji Zhang. Towards characterizing markov equivalence classes for directed acyclic graphs with latent variables. In *21st Conference on Uncertainty in Artificial Intelligence, UAI 2005*, 2005. 5

[AS04] Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2004. 17

[BFR+00] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '00, pages 259–269, Washington, DC, USA, 2000. IEEE Computer Society. 6

[BFRV11] Arnab Bhattacharyya, Eldar Fischer, Ronitt Rubinfeld, and Paul Valiant. Testing monotonicity of distributions over general partial orders. In *ICS*, pages 239–252, 2011. 6

[BL92] Kenneth A Bollen and J Scott Long. Tests for structural equation models: introduction. *Sociological Methods & Research*, 21(2):123–131, 1992. 4

[BMS08] Guy Bresler, Elchanan Mossel, and Allan Sly. Reconstruction of markov random fields from samples: Some observations and algorithms. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 343–356. Springer, 2008. 6

[BP12] Elias Bareinboim and Judea Pearl. Transportability of causal effects: Completeness results. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 698–704. AAAI Press, 2012. 5

[BP13] E. Bareinboim and J. Pearl. Meta-transportability of causal effects: A formal approach. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 135–143, 2013. 5

[Bre15] Guy Bresler. Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the 47th Annual ACM Symposium on the Theory of Computing*, STOC '15, pages 771–782, New York, NY, USA, 2015. ACM. 6

[Can15] Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22(63), 2015. 6

[CDKS17] Clement L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing bayesian networks. In *Conference on Learning Theory*, pages 370–448, 2017. 4, 6, 8

[CL68] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968. 6

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006. 6

[DDK17] Constantinos Daskalakis, Nishanth Dikkala, and Gautam C. Kamath. Concentration of Multilinear Functions of the Ising Model with Applications to Network Data. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 6

[DDK18] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing Ising models. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, Philadelphia, PA, USA, 2018. SIAM. 6

[DK16] Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. *CoRR*, abs/1601.05557, 2016. 6, 9

[DL12] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012. 9

[DP17] Constantinos Daskalakis and Qinxuan Pan. Square hellinger subadditivity for bayesian networks and its applications to identity testing. *Proceedings of Machine Learning Research vol*, 65:1–7, 2017. 4, 6, 7, 8

[DPL+16] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychndhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-seq: Dissecting molecular circuits with scalable single cell rna profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, Dec 2016. 27984732[pmid]. 2, 5

[Ebe07] Frederick Eberhardt. Causation and intervention. *Doctoral dissertation, Carnegie Mellon University*, 2007. 5

[EGL+92] Guy Even, Oded Goldreich, Michael Luby, Noam Nisan, and Boban Veličkovic. Approximations of general independent distributions. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, pages 10–16. ACM, 1992. 16

[EGS05] Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 178–184. AUAI Press, 2005. 5

[Fis25] Ronald Aylmer Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925. 6

[GC99] Clark N Glymour and Gregory Floyd Cooper. *Computation, causation, and discovery*. AAAI Press, 1999. 2

[GJS10] Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf. Causality: Objectives and assessment. In *Causality: Objectives and Assessment*, pages 1–42, 2010. 5

[Gol17] Oded Goldreich. *Introduction to property testing*. Cambridge University Press, 2017. 6

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(20), 2000. 6

[Haa43] Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, pages 1–12, 1943. 2

[HB12a] Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.*, 13(1):2409–2464, August 2012. 5

[HB12b] Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal networks from interventional data. In *Proceedings of Sixth European Workshop on Probabilistic Graphical Models*, volume 119, 2012. 5

[HEH13] Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Experiment selection for causal discovery. *The Journal of Machine Learning Research*, 14(1):3041–3071, 2013. 5

[HJM+09] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009. 4

[JMZ+12] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012. 5

[KDV17] Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. Cost-optimal learning of causal graphs. In *International Conference on Machine Learning*, pages 1875–1884, 2017. 5

[KF09] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 2

[KM17] Adam Klivans and Raghu Meka. Learning graphical models using multiplicative weights. *arXiv preprint arXiv:1706.06274*, 2017. 6

[KSB17] Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. In *Advances in Neural Information Processing Systems*, pages 7021–7031, 2017. 5

[KT06] Changsung Kang and Jin Tian. Inequality constraints in causal models with hidden variables. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 233–240. AUAI Press, 2006. 5

[LH13] Sanghack Lee and Vasant Honavar. m-transportability: Transportability of a causal effect from multiple environments. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA.*, 2013. 5

[LR06] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006. 6

[MBS+15] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, May 2015. 26000488[pmid]. 5

[MMLM06] Stijn Meganck, Sam Maes, Philippe Leray, and Bernard Manderick. Learning semi-markovian causal models using experiments. In *Proceedings of The third European Workshop on Probabilistic Graphical Models (PGM)*, 2006. 5

[Mos09] Robin A Moser. A constructive proof of the lovász local lemma. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 343–350. ACM, 2009. 17

[MT10] Robin A Moser and Gábor Tardos. A constructive proof of the general lovász local lemma. *Journal of the ACM (JACM)*, 57(2):11, 2010. 17

[Nea04] Richard E Neapolitan. *Learning bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004. 2

[PB11] J. Pearl and E. Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*, pages 247–254, Menlo Park, CA, August 7-11 2011. Available at: <http://ftp.cs.ucla.edu/pub/stat_ser/r372a.pdf>. 5

[Pea95] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. 10

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009. 1, 2, 3, 8

[PJS11] Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2436–2450, 2011. 4

[PV95] Judea Pearl and Thomas S Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995. 5

[SDLC93] David J Spiegelhalter, A Philip Dawid, Steffen L Lauritzen, and Robert G Cowell. Bayesian analysis in expert systems. *Statistical science*, pages 219–247, 1993. 2

[SGS00] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000. 2, 5

[SKDV15] Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. In *Advances in Neural Information Processing Systems*, pages 3195–3203, 2015. 5

[SMR99] P Spirtes, C Meek, and T Richardson. An algorithm for causal inference in the presence of latent variables and selection bias in computation, causation and discovery, 1999, 1999. 5

[SP08] I. Shpitser and J. Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008. 5

[SPP+05] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005. 5

[SS08] Richard Scheines and Peter Spirtes. Causal structure search: Philosophical foundations and problems, 2008. 5

[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. 8

[SSS+17] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*, pages 2955–2965, 2017. 4

[SW12] Narayana P Santhanam and Martin J Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012. 6

[Tia02] Jin Tian. *Studies in causal reasoning and learning*. University of California, Los Angeles, 2002. 8

[TP02] 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. 3, 5, 6, 11, 17, 34, 35

[VMLC16] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2016. 6

[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. 6, 34, 35

[VP92] Thomas Verma and Judea Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Uncertainty in Artificial Intelligence, 1992*, pages 323–330. Elsevier, 1992. 5

[Wri21] Sewall Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921. 2

[WSYU17] Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5822–5831. Curran Associates, Inc., 2017. 5

[YKU18] Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. *arXiv preprint arXiv:1802.06310*, 2018. 5

[Zha08] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008. 5

[ZPJS12] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012. 4

A Proof Sketch for the Fully Observable Case

In the absence of unobservable variables, the analysis becomes much simpler. Let us look at the two-sample testing problem on input causal models $\mathcal{X}$ and $\mathcal{Y}$ defined on a DAG G. Now, each c-component is a single vertex, so that every "local" intervention is of the form $P[V_i \mid do(\mathbf{pa}(V_i))]$ for a vertex V_i and an assignment $\mathbf{pa}(V_i)$ to the parents of V_i. We define our tester to accept iff each such local intervention on $\mathcal{X}$ and $\mathcal{Y}$ yields distributions which differ by at most $\varepsilon^2/2n$ in squared Hellinger distance. The squared Hellinger distance is defined as follows for two distributions P and Q on $[D]$:

$$H^2(P,Q) := 1 - \sum_{i\in[D]} \sqrt{P(i)\cdot Q(i)} = 1 - BC(P,Q) \tag{9}$$

where $BC(P,Q)$ is the Fidelity or Bhattacharya coefficient of P and Q. Below, our subadditivity theorem shows that if the algorithm accepts, then for every intervention, the resulting distributions for $\mathcal{X}$ and $\mathcal{Y}$ differ by at most $\varepsilon^2/2$ in squared Hellinger distance, implying $\Delta(\mathcal{X},\mathcal{Y}) \leqslant \varepsilon$.

Theorem A.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be two causal Bayesian networks defined on a* known and common *DAG G with no hidden variables. Identify the vertices in $\mathbf{V}$ as $\{V_1,\ldots,V_n\}$ arranged in a topological order. Suppose we know that*

$$H^2(P_{\mathcal{X}}[V_j \mid do(\mathbf{pa}(V_j))], P_{\mathcal{Y}}[V_j \mid do(\mathbf{pa}(V_j))]) \leqslant \gamma \qquad \forall j \in [n], \forall\, \mathbf{pa}(V_j) \in \Sigma^{|\,\mathbf{Pa}(V_j)|}. \tag{10}$$

Then, for each subset $\mathbf{T} \subseteq \mathbf{V}$ and $\mathbf{t} \in \Sigma^{|\mathbf{T}|}$,

$$H^2\left(P_{\mathcal{X}}[\mathbf{V}\setminus\mathbf{T} \mid do(\mathbf{t})], P_{\mathcal{Y}}[\mathbf{V}\setminus\mathbf{T} \mid do(\mathbf{t})]\right) \leqslant \gamma n. \tag{11}$$

Proof. Fix $\mathbf{T} \subseteq \mathbf{V}$ and an assignment $\mathbf{t} \in \Sigma^{|\mathbf{T}|}$. Let $\mathbf{W} = \mathbf{V}\setminus\mathbf{T} = \{W_1, W_2, \ldots, W_m\}$ whose indices are arranged in a topological ordering. By the definition of squared Hellinger distance:

$$\begin{aligned}
&H^2\begin{pmatrix} P_{\mathcal{X}}[\mathbf{W}|do(\mathbf{t})], \\ P_{\mathcal{Y}}[\mathbf{W}|do(\mathbf{t})] \end{pmatrix} \\
&= 1 - \sum_{w_1,w_2,\ldots,w_m} \sqrt{\begin{matrix} P_{\mathcal{X}}[w_1,w_2,\ldots,w_m|do(\mathbf{t})] \\ P_{\mathcal{Y}}[w_1,w_2,\ldots,y_m|do(\mathbf{t})] \end{matrix}} \\
&= 1 - \sum_{w_1,\ldots,w_{m-1}} \sqrt{\begin{matrix} P_{\mathcal{X}}[w_1,\ldots,w_{m-1}|do(\mathbf{t})] \\ P_{\mathcal{Y}}[w_1,\ldots,w_{m-1}|do(\mathbf{t})] \end{matrix}} \sum_{w_m} \sqrt{\begin{matrix} P_{\mathcal{X}}[w_m|w_1,\ldots,w_{m-1},do(\mathbf{t})] \\ P_{\mathcal{Y}}[w_m|w_1,\ldots,w_{m-1},do(\mathbf{t})] \end{matrix}} \\
&= 1 - \sum_{w_1,\ldots,w_{m-1}} \sqrt{\begin{matrix} P_{\mathcal{X}}[w_1,\ldots,w_{m-1}|do(\mathbf{t})] \\ P_{\mathcal{Y}}[w_1,\ldots,w_{m-1}|do(\mathbf{t})] \end{matrix}} \sum_{w_m} \sqrt{\begin{matrix} P_{\mathcal{X}}[w_m|do(\mathbf{pa}(w_m))] \\ P_{\mathcal{Y}}[w_m|do(\mathbf{pa}(w_m))] \end{matrix}}.
\end{aligned}$$

The above step can be obtained easily by using Lemma C.1 and the conditional independence constraints obtained from G. Therefore:

$$\begin{aligned}
H^2\begin{pmatrix} P_{\mathcal{X}}[\mathbf{W}|do(\mathbf{t})], \\ P_{\mathcal{Y}}[\mathbf{W}|do(\mathbf{t})] \end{pmatrix} &\leqslant 1 - \sum_{w_1,\ldots,w_{m-1}} \sqrt{\begin{matrix} P_{\mathcal{X}}[w_1,\ldots,w_{m-1}|do(\mathbf{t})] \\ P_{\mathcal{Y}}[w_1,\ldots,w_{m-1}|do(\mathbf{t})] \end{matrix}}\,(1-\gamma) \quad \text{(from (10))} \\
&= H^2\begin{pmatrix} P_{\mathcal{X}}[W_1\ldots W_{m-1} \mid do(\mathbf{t})], \\ P_{\mathcal{Y}}[W_1\ldots W_{m-1}|do(\mathbf{t})] \end{pmatrix}(1-\gamma) + \gamma.
\end{aligned}$$

By induction on n, we get:

$$H^2\left(\begin{array}{c} P_{\mathcal{X}}[\mathbf{W}|do(\mathbf{t})], \\ P_{\mathcal{Y}}[\mathbf{W}|do(\mathbf{t})] \end{array}\right) \leqslant \gamma[1+(1-\gamma)+(1-\gamma)^2+\ldots+(1-\gamma)^{m-1}]$$
$$= 1-(1-\gamma)^m \leqslant 1-(1-\gamma)^n \leqslant n\gamma.$$

□

The time and sample complexities are then determined by that required for two-sample testing on each pair of local distributions with accuracy $\varepsilon^2/2n$ in H^2 distance. We defer this calculation, as well as bounding the total number of interventions, to later when we analyze semi-Markovian CBNs.

B Reduction from General Graphs

First we define the effective parents and the c-component relation for general causal graphs.

Definition B.1 (Effective Parents $\mathbf{Pa}^+$)**.** *Given a general causal graph H and a vertex $V_i \in \mathbf{V}$,* the effective parents of V_i*, denoted by $\mathbf{Pa}^+(V_i)$, is the set of all observable vertices V_j such that either V_j is a parent of V_i or there exists a directed path from V_j to V_i that contains only unobservable variables.*

Definition B.2 (c-component)**.** *For a given general causal graph H, two vertices V_i and V_j are related by the* c-component relation *if (i) there exists an unobservable variable U_k such that H contains two paths (i) from U_k to V_i; and (ii) from U_k to V_i, where both the paths use only unobservable variables, or (ii) there exists another vertex $V_z \in \mathbf{V}$ such that V_i and V_z (and) V_j and V_z are related by c-component relation.*

We study Semi Markovian Bayesian Networks (SMBN)'s without any loss of generality owing to the projection of a general causal graph to a SMCG [TP02, VP90]. For a given graph H they showed that there is an equivalent SMCG G such that the c-component factorization and some other important properties hold. Namely,

- The set of observable nodes in H and G are the same.
- The topological ordering of the observable nodes in H and G are the same.
- The c-components of H and G are identical and the c-component factorization formula (Lemma 2.12 here, (20) in Lemma 2 of [TP02]) holds even for the general causal graph (See Section 5 of [TP02]). They show this based on a known previously known reduction from H to G [VP90]. The proof is based on the fact that for any subset $\mathbf{S} \subseteq \mathbf{V}$ of observable variables, the induced subgraphs $G[\mathbf{S}]$ and $H[\mathbf{S}]$ require the same set of conditional independence constraints.
- The parents of nodes in G are the effective parents of nodes in H.

All the results presented in this paper depend only on the above mentioned properties. Therefore, we can reduce the given general causal graph H to a SMCG G using the available reduction and work with G, where the parents of vertices of G correspond to the effective parents of the respectives vertices of H. Now we proceed to show the algorithm of [VP90] that preserves all the required properties mentioned above.

Projection Algorithm of [TP02, VP90] For a given causal graph H, the projection algorithm reduces the given causal graph H to a SMCG G by the following procedure:

1. For each observable variable $V_i \in V$ of H, add an observable variable V_i in G.
2. For each pair of observable variables $V_i, V_j \in \mathbf{V}$, if there exists a directed edge from V_i to V_j in H, or if there exists a *directed* path from V_i to V_j that contains only unobservable variables in H, then add a directed edge from V_i to V_j in G.
3. For each pair of observable variables $V_i, V_j \in \mathbf{V}$, if there exists an unobservable variable U_k such that there exist two *directed* paths in H from U_k to V_i and from U_k to V_j such that both the paths contain only the unobservable variables, then add a bi-directed edge between V_i and V_j in G.

C Conditional Independence

The following lemma captures a useful fact about conditional independence between variables in a SMBN.

Lemma C.1 (Independence Lemma)**.** *Let M be a* SMBN *with respect to a* SMCG *G with the vertex set $\mathbf{V} = \{V_1, \ldots, V_n\}$ (where the indices respect topological ordering). For a given intervention $do(\mathbf{t})$, let $\mathbf{C} = \{V_{n_1}, V_{n_2}, \ldots, V_{n_s}\}$ be a c-component of the induced subgraph $G' = G[\mathbf{V} \setminus \mathbf{T}]$, where $s = |\mathbf{C}|$ and $n_1 < n_2 < \cdots < n_s$. Then for a given vertex V_{n_i}, for a given set $\mathbf{D}$ such that $\mathbf{V} \setminus (\mathbf{T} \cup \{V_{n_1}, \ldots, V_{n_i}\}) \supseteq \mathbf{D} \supseteq \mathbf{Pa}_{G'}(\{V_{n_1}, \ldots, V_{n_i}\})$, and a given set of assignments $v_{n_1}, \ldots, v_{n_i}$, $\mathbf{d}$,*

$$P_{\mathcal{M}}[v_{n_i} \mid v_{n_1}, \ldots, v_{n_{i-1}}, do(\mathbf{d}, \mathbf{t})] = P_{\mathcal{M}}[v_{n_i} \mid v_{n_1}, \ldots, v_{n_{i-1}}, do(\mathbf{pa}_{G'}(V_{n_1}, \ldots, V_{n_i}), \mathbf{t})]$$

where $pa_{G'}(v_{n_1}, \ldots, v_{n_i})$ is the assignment that is consistent with $\mathbf{D}$.

Proof. By Bayes' theorem

$$P_{\mathcal{M}}\left[v_{n_{j,i}} \middle| \begin{array}{l} v_{n_{j,1}}, \ldots, v_{n_{j,i-1}}, \\ do(\mathbf{pa}_{G'}(V_{n_{j,1}}, \ldots, V_{n_{j,i}}), \mathbf{t}) \end{array}\right] = \frac{P_{\mathcal{M}}[v_{n_{j,i}}, v_{n_{j,1}}, \ldots, v_{n_{j,i-1}} \mid do(\mathbf{pa}_{G'}(V_{n_{j,1}}, \ldots, V_{n_{j,i}}), \mathbf{t})]}{P_{\mathcal{M}}[v_{n_{j,1}}, \ldots, v_{n_{j,i-1}} \mid do(\mathbf{pa}_{G'}(V_{n_{j,1}}, \ldots, V_{n_{j,i}}), \mathbf{t})]}. \tag{12}$$

We apply Lemma 2.11 with respect to the graph $G' = G[\mathbf{V} \setminus \mathbf{T}]$ that is obtained after the intervention $do(\mathbf{t})$ for both the numerator and the denominator of (12) seperately. Therefore:

$$\begin{aligned} P_{\mathcal{M}}\left[v_{n_{j,i}} \middle| \begin{array}{l} v_{n_{j,1}}, \ldots, v_{n_{j,i-1}}, \\ do(\mathbf{pa}_{G'}(V_{n_{j,1}}, \ldots, V_{n_{j,i}}), \mathbf{t}) \end{array}\right] &= \frac{P_{\mathcal{M}}[v_{n_{j,i}}, v_{n_{j,1}}, \ldots, v_{n_{j,i-1}} \mid do(\mathbf{d}, \mathbf{t})]}{P_{\mathcal{M}}[v_{n_{j,1}}, \ldots, v_{n_{j,i-1}} \mid do(\mathbf{d}, \mathbf{t})]} \\ &= P_{\mathcal{M}}[v_{n_{j,i}} \mid v_{n_{j,1}}, \ldots, v_{n_{j,i-1}}, do(\mathbf{d}, \mathbf{t})]. \end{aligned}$$

□