Proceedings of Machine Learning Research vol 117:1–29, 2020 31st International Conference on Algorithmic Learning Theory

On Learning Causal Structures from Non-Experimental Data without Any Faithfulness Assumption

Hanti Lin IKA@UCDAVIS.EDU
Philosophy Department, University of California, Davis

Jiji Zhang JIJIZHANG@LN.EDU.HK
Philosophy Department, Lingnan University

Editors: Aryeh Kontorovich and Gergely Neu

Abstract

Consider the problem of learning, from non-experimental data, the causal (Markov equivalence) structure of the true, unknown causal Bayesian network (CBN) on a given, fixed set of (categorical) variables. This learning problem is known to be very hard, so much so that there is no learning algorithm that converges to the truth for all possible CBNs (on the given set of variables). So the convergence property has to be sacrificed for some CBNs—but for which? In response, the standard practice has been to design and employ learning algorithms that secure the convergence property for at least all the CBNs that satisfy the famous *faithfulness* condition, which implies sacrificing the convergence property for some CBNs that violate the faithfulness condition (Spirtes, Glymour, and Scheines, 2000). This standard design practice can be justified by assuming—that is, accepting on faith—that the true, unknown CBN satisfies the faithfulness condition. But the real question is this: Is it possible to explain, *without assuming* the faithfulness condition or any of its weaker variants, why it is mandatory rather than optional to follow the standard design practice? This paper aims to answer the above question in the affirmative. We first define an array of modes of convergence to the truth as desiderata that might or might not be achieved by a causal learning algorithm. Those modes of convergence concern (i) how pervasive the domain of convergence is on the space of all possible CBNs and (ii) how uniformly the convergence happens. Then we prove a result to the following effect: for *any* learning algorithm that tackles the causal learning problem in question, if it achieves the best achievable mode of convergence (considered in this paper), then it *must* follow the standard design practice of converging to the truth for at least all CBNs that satisfy the faithfulness condition—it is a requirement, not an option.

Keywords: Causal Bayesian Network, Causal Discovery, Faithfulness Condition, Learning Theory, Almost Everywhere Convergence, Locally Uniform Convergence

1. Introduction

Suppose that there is a causal system that can be properly modeled by some causal Bayesian network (CBN) on a given set of observable variables, and that we aim to learn the causal structure of the true, unknown CBN (at least up to Markov equivalence), which is crucial to predicting what causal effects there would be if we were to manipulate this or that variable. Suppose, further, that we wish to learn the causal structure only from non-experimental data, possibly because experimentation is too costly or unethical. It is well-known that this learning problem is very hard. The difficulty is that there can be two very different CBNs that are indistinguishable in terms of non-experimental data. To be more precise, there can be two CBNs N and N' with the following properties:

1. (CAUSAL DIFFERENCE) N and N' have quite *different* causal structures that are not even Markov equivalent.

2. (STATISTICAL NONIDENTIFIABILITY) N and N' share the *same* joint probability distribution; so, if a learning algorithm receives only non-experimental data (i.e., data collected without causing a change in the joint distribution), then it must, at any sample size, fail to have a high probability of identifying the true structure either for N or for N'.

Because of property 1 (causal difference), it would be great if we could have a learning algorithm that converges in probability to the true causal structure (up to Markov equivalence) for both N and N' and, hopefully, for all CBNs on the given set of variables. But, unfortunately, no learning algorithm can be that good, by property 2 (statistical nonidentifiability). So, when we design a causal learning algorithm (also called causal discovery algorithm), the convergence property *must be sacrificed* for N or for N' and similarly for any other pair of CBNs to the same effect. Sacrifices have to be made for some—but for which?

In reaction to that difficulty, the standard practice has been to design and employ learning algorithms that secure the convergence property for *at least* all the CBNs that satisfy the famous *faithfulness* condition (Spirtes, Glymour, and Scheines, 2000), which implies sacrificing the convergence property for some (possibly not all) CBNs that violate the faithfulness condition. Examples abound, including constraint-based algorithms such as *PC* (Spirtes et al., 2000), score-based algorithms such as *GES* (Chickering, 2002), and hybrids of those two kinds of algorithms (Zhalama et al., 2017). This standard design practice can be justified if we are willing to simply assume—that is, accept on faith—that the unknown, true CBN turns out to satisfy the faithfulness condition. But the real question is this: *Can we justify this standard design practice without assuming the faithfulness condition or any of its weaker variants?* This is the question that this paper aims to address—and answer in the affirmative.

While we believe that the question just posed is very important, there is only a very small literature that attempts to address it. As far as we know, very few works try to address the issue explicitly; two notable ones are Spirtes et al. (2000) and Meek (1995). They show that the faithfulness condition only rules out a mathematically negligible set of CBNs (in the sense of negligibility that, roughly, a lower-dimensional plane is negligible in a higher-dimensional space). So any learning algorithm that follows the standard design practice sacrifices the convergence property only for a negligible set of CBNs. So it seems that we should not worry too much about using such a learning algorithm.

But some question remains to be addressed. To be sure, we should not worry about using a learning method that sacrifices the convergence property for some CBNs that form a mathematically negligible set, precisely because no learning method can avoid sacrificing that much. So sacrifices have to be made at least for *some* mathematically negligible set of CBNs—but for *which* should sacrifices be made? One option is to follow the standard design practice. But there are alternatives, such as this one: (i) identify two CBNs N and N' that share the same joint distribution, with N satisfying the faithfulness condition and N' violating that condition, (ii) design and adopt a learning algorithm whose domain of convergence to the truth is the same as the set of faithful CBNs except that the faithful one N is removed from the domain of convergence and the unfaithful one N' is added to it. Such a learning algorithm is one of the infinitely many alternatives that sacrifice the convergence property only for a mathematically negligible set of CBNs but run counter to the standard design practice, which tries to secure the convergence property for at least *all* faithful

CBNs. *But should we follow the standard design practice rather than any of those competing alternatives? If so, why?* That's the question.

To answer the above question, this paper develops a general, straightforward strategy:

> GENERAL STRATEGY. When the learning problem in question is extremely hard, so much so that (almost) every familiar desideratum for learning algorithms is provably too high an ideal to be achievable, we do not have to react by making an assumption that turns the learning problem into an easier one. Instead, we can react in this way: keep the learning problem as it is, look for what can be achieved, and determine what it takes for a learning method to achieve the highest achievable desideratum.

This strategy is implemented by defining certain modes of convergence to the truth that can be taken as desiderata for learning algorithms. To begin with, consider the question of where convergence happens. It would be great to extend the domain of convergence to cover everywhere on the space of all CBNs (on the given set of variables); but that is provably impossible for causal learning. So we examine the possibility of achieving some lower ideals (and their combinations):

(a) extending the domain of convergence to cover *almost everywhere* on the space of all CBNs, i.e., everywhere except on a topologically negligible subset (a nowhere dense subset);

(b) having a *maximal* domain of convergence, i.e., one that cannot be extended further.

This leads to the modes of convergence listed on the axis in figure 1 that stretches to the upper right. The other axis, which stretches to the upper left, concerns the question of how uniformly convergence happens. It would be great to achieve globally uniform convergence (aka uniform

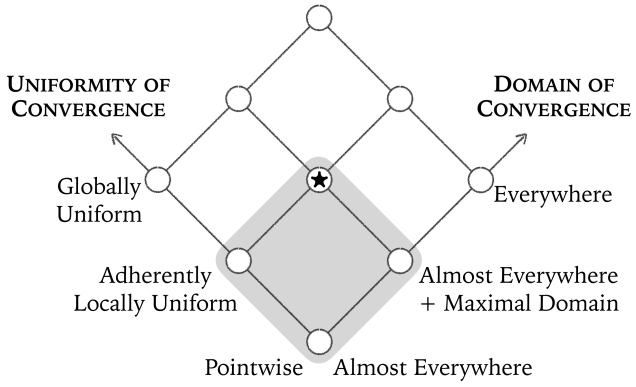


Figure 1: Modes of convergence to the truth

consistency), but that is provably impossible for causal learning. That is, no causal learning algorithm can guarantee a bound on the error probability that applies to all CBNs. So we examine the possibility of achieving something local rather than global:

(c) *locally uniform* convergence of a certain kind—the "adherent" kind—that guarantees that a low error probability can be obtained stably under small perturbations of the joint probability distribution but without a change in the causal structure (that is, with "adherence" to the unchanged causal structure).

Those two considerations—about domain of convergence and uniformity of convergence—are then used to define nine joint modes of convergence, depicted as the nine nodes in the lattice in figure 1. (Some of those joint modes are equivalent because globally uniform convergence, alone, is strong enough to imply anything else in the figure.) The main result of this paper is theorem 1 (stated in section 4), which can be summarized as follows:

- Of the joint modes of convergence in figure 1, the achievable ones for the problem of learning the true causal structure (up Markov equivalence) are exactly those in the shaded area. So the best achievable one is the one marked by a star $\star$ in the figure, which conjoins the three modes of convergence (a) "almost everywhere", (b) "maximal domain", and (c) "adherently locally uniform".

- For *any* causal learning algorithm L, if L satisfies at least (a), (b), and (c) simultaneously (whether or not it satisfies any additional desiderata about, say, rates of convergence or computational complexity), then L *must* follow the standard design practice in that it converges to the truth for at least all CBNs that satisfy the faithfulness condition—this is a requirement, not an option.

To the best of our knowledge, this is the first theoretical result that explains, without assuming the faithfulness condition or any of its weaker variants, why it is mandatory rather than optional to follow the standard design practice. This result is proved for any fixed finite set of categorical variables, under just the standard assumption of IID (identically and independently distributed observations) and the assumptions built into the definition of causal Bayesian networks.

This paper actually does more. To achieve at least the desideratum of (a)+(b)+(c), convergence to the truth *must be secured* for some range of CBNs, *must be sacrificed* for some other range, and is *optional* for the remaining range. Those three ranges are precisely determined in a strengthening of the main result, theorem 2 (stated in section 5).

The rest of this paper proceeds as follows. Standard definitions are reviewed in sections 2 with examples. Key definitions are provided and motivated in section 3. The main result is stated and discussed in section 4, followed by a strengthened result in section 5. Section 6 provides an illustrated proof of a quite revealing lemma. Complete proofs are left to the appendix. To declare the style in use: Emphasis is indicated by *italics*; the terms to be defined are presented in **boldface**.

2. Review of Standard Definitions

Fix a finite set of variables, $\mathcal{V} = \{X_1, X_2, \ldots, X_K\}$. A possible **causal structure** over those variables is represented by a directed acyclic graph on $\mathcal{V}$, written $G = (\mathcal{V}, \rightarrow)$, where the binary relation $X_i \rightarrow X_j$ is understood to say that X_i is an immediate cause of X_j relative to $\mathcal{V}$, or in short, that X_i is a **parent** of X_j. If a variable X_i is a parent of (a parent of a parent of ...) a variable X_j, say that X_j is a **descendant** of X_i. For convenience, we count every variable as its own descendant. We will refer to G simply as a **(causal) graph**, dropping 'directed acyclic', because only directed acyclic graphs are considered in this paper. If a graph G and a joint distribution P are so connected that each variable in G is P-independent of its non-descendants given all of its parents (with respect to graph G), say that graph G and distribution P satisfy the **Markov condition**, that G is **Markov** to P, and that (G, P) is a **causal Bayesian network** (CBN). The Markov condition, as the defining condition of CBNs, is often taken for granted as a necessary connection between causal graphs and joint distributions—between causation and probability.

The Markov condition can be conveniently expressed in another way. Let $\mathcal{V}_1$, $\mathcal{V}_2$, and $\mathcal{V}_3$ be disjoint subsets of the given, fixed set $\mathcal{V}$ of variables. Understand $\mathcal{V}_1 \perp\!\!\!\perp \mathcal{V}_2 \mid \mathcal{V}_3$ as the statement saying that $\mathcal{V}_1$ and $\mathcal{V}_2$ are independent given $\mathcal{V}_3$. A graph G is said to **entail** a conditional independence statement $\mathcal{V}_1 \perp\!\!\!\perp \mathcal{V}_2 \mid \mathcal{V}_3$ if that statement holds with respect to every joint distribution to which G is Markov. Let $\mathcal{I}(G)$ denote the set of the conditional independence statements that G entails. Let $\mathcal{I}(P)$ denote the set of the conditional independence statements that hold with respect to P. Then it is well-known that G and P satisfy the Markov condition if and only if

$$\mathcal{I}(G) \subseteq \mathcal{I}(P)\,.$$

Now, consider the following stronger condition:

$$\mathcal{I}(G) = \mathcal{I}(P)\,.$$

This condition says that the conditional independence statements entailed by G are exactly those that hold with respect to P; in that case, say that G is **faithful** to P, and that the CBN (G, P) satisfies the **faithfulness condition**. With respect to an **unfaithful** CBN (G, P), at least one conditional independence statement σ turns out to hold (i.e., $\sigma \in \mathcal{I}(P)$) even though it is not required to hold by the Markov condition (i.e., $\sigma \notin \mathcal{I}(G)$).

$P(X=1)$
0.5

x	$P(Y=1 \mid X=x)$
1	0.2
0	0.8

x	y	$P(Z=1 \mid X=x, Y=y)$
1	1	p_1
1	0	p_2
0	1	p_3
0	0	0

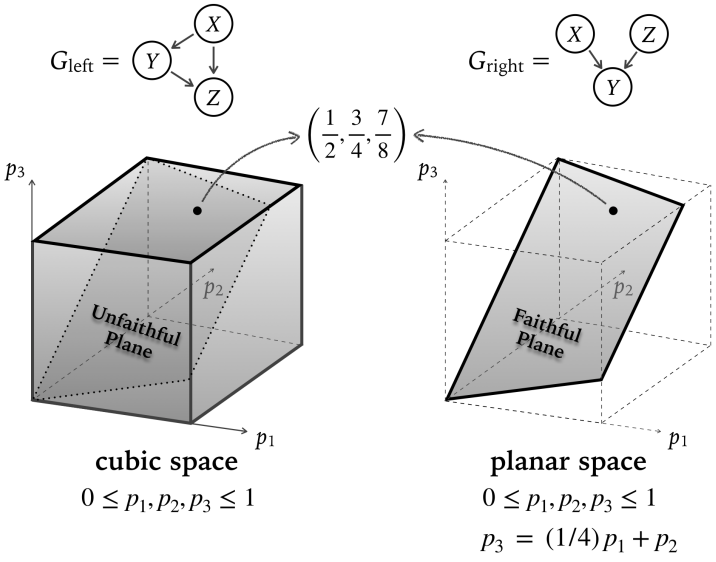


Figure 2: Two spaces of causal Bayesian networks

To illustrate, let $\mathcal{V}$ contain only three binary variables X, Y, and Z. Consider the two causal graphs depicted in figure 2, G_{left} and G_{right}. Instead of thinking about all joint distributions on $\mathcal{V}$, for the sake of visualization let's consider just the joint distributions that are defined by the three tables on the left of the same figure, with three parameters p_1, p_2, and p_3 taking values in the unit interval. So those parameterized distributions form a unit cube. The design of this parameterized family ensures that the left causal graph G_{left} is Markov to each of those distributions. So the points in the left cube in the same figure represent all the CBNs (G_{left}, P) with P in the parameterized family.

The right graph G_{right}, on the other hand, turns out to be Markov *only* to the distributions (p_1, p_2, p_3) under the constraint $p_3 = (1/4)p_1 + p_2$, which defines the trapezoidal plane on the right in figure 2. Here is why. The left graph G_{left} entails only the conditional independence statements that are trivially true (i.e., true simply in virtue of the probability calculus). But the right graph G_{right} entails one more conditional independence statement: $\{X\} \perp\!\!\!\perp \{Z\} \mid \varnothing$, which says that X and Z are independent. To satisfy this additional independence relation is provably to satisfy the equation $p_3 = (1/4)p_1 + p_2$ that defines the trapezoidal plane on the right.

The cubic space of CBNs on the left embeds a copy of the trapezoidal plane, as depicted in the same figure. So every CBN (G_{left}, P) on the left, embedded trapezoid shares the same joint distribution P with its counterpart CBN (G_{right}, P) on the right. Due to this sharing of the same joint distribution, every CBN (G_{left}, P) on the left, embedded trapezoid is unfaithful because it satisfies the independence between X and Z, which goes beyond the conditional independence statements entailed by the left causal graph G_{left}. So the left, embedded trapezoid contains only unfaithful CBNs; accordingly, call it the **unfaithful plane**. The sharing of the same joint distribution is an important source of the difficulty of causal learning, as we will see below.

A **causal learning problem** is represented by a triple $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ whose three components are understood as follows:

1. (VARIABLE) A certain causal system is under study and assumed or known to be accurately represented by a CBN on a given set $\mathcal{V}$ of variables.

2. (STATE) The true, unknown CBN is assumed, and only assumed, to be in a given set $\mathcal{S}$ of CBNs over those variables. Each element of $\mathcal{S}$ is a CBN (G, P) understood as a possible state of the causal world, called a **causal state**, in which the true joint probability distribution is P and the true causal structure is G. So $\mathcal{S}$ is the space of the possible causal states under consideration.

3. (HYPOTHESIS) The goal is to learn, from non-experimental data, the truth among certain competing hypotheses that form a given set $\mathcal{H}$—hypotheses about the causal structure of the true, unknown CBNs. In every causal state in $\mathcal{S}$, exactly one causal hypothesis in $\mathcal{H}$ is true.

Given a causal learning problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$, a learning method for tackling that problem is, roughly, a function that maps each possible "data set" to a hypothesis in $\mathcal{H}$; such a learning method might perform "well" in one causal state but "poorly" in another, and will be evaluated in terms of its (varying) performances in all the causal states in the given state space $\mathcal{S}$. Those rough ideas can be made precise as follows.

To define data sets, assume for simplicity that we can observe the value of every variable in $\mathcal{V}$. Fix an enumeration of the variables $X_1, X_2, \ldots, X_K$ in $\mathcal{V}$. Let $\mathbf{X} = (X_1, X_2, \ldots, X_K)^{\mathrm{T}}$ be the column vector of those variables, in the order of the given enumeration. The observation of the i-th instance of the causal system under study will be represented by a random vector $\mathbf{X}_i = (X_{1,i}, X_{2,i}, \ldots, X_{K,i})^{\mathrm{T}}$, where $X_{k,i}$ represents the observation of the k-th variable X_k in the i-th instance of the causal system. Assume that observations $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_i, \ldots$ of different instances are independent and identically distributed (IID). So, if (G, P) is the true CBN, then $\mathbf{X}_i \sim \mathbf{X} \sim P$ for each $i \geq 1$. A **data set** of sample size n (over the set $\mathcal{V}$ of variables), written $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, is a realization of the n observations $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$. The collection of all such data sets, written $\mathsf{Data}(\mathcal{V})$, contains the possible inputs considered in this paper.

A **learning method** for tackling a causal learning problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ is formally a function from $\mathsf{Data}(\mathcal{V})$ to $\mathcal{H}$—a function $\hat{H}$ that maps any data set $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of any sample size n (over the set $\mathcal{V}$ of variables) to a hypothesis in $\mathcal{H}$, denoted by $\hat{H}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. To clarify, $\hat{H}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is a *specific* causal hypothesis, one that the learning method $\hat{H}$ outputs/accepts when it receives a concrete data set $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. In contrast, $\hat{H}(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is a *random* causal hypothesis—a random variable denoting the causal hypothesis that $\hat{H}$ outputs given a random observation $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ of sample size n. While a learning method is construed abstractly as a mere input-output relation, a **learning algorithm** is understood as a concrete instruction that (efficiently or inefficiently) implements some learning method. This paper focuses on the abstract level of learning methods rather than the concrete level of learning algorithms.

In a causal state $s = (G, P)$, a learning method $\hat{H}$ might perform well or poorly given a sample size n, and the performance is captured by the following probabilities:

$$\begin{aligned} \textbf{success probability} &= \mathbb{P}_s\left(\hat{H}(\mathbf{X}_1, \ldots, \mathbf{X}_n) = H_s\right), \\ \textbf{error probability} &= \mathbb{P}_s\left(\hat{H}(\mathbf{X}_1, \ldots, \mathbf{X}_n) \neq H_s\right), \end{aligned}$$

where $\mathbb{P}_s$ denotes the sampling distribution true in causal state $s = (G, P)$, namely the ∞-fold probability measure generated by P under the IID assumption, and H_s denotes the causal hypothesis in $\mathcal{H}$ that is true in causal state $s = (G, P)$.

3. Key Definitions: Modes of Convergence

The crux of the matter can be understood this way: When the learning problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ in question makes too weak a background assumption—that is, when the state space $\mathcal{S}$ under consideration is too inclusive—then (almost) all familiar desiderata for learning methods become too high an ideal to be achievable. For example, consider the familiar desideratum of *statistical consistency*, which can be defined in the present setting as follows. A learning method $\hat{H}$ for tackling a causal learning problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ is said to **converge (in probability) to the truth** in a causal state $s = (G, P)$ if, in state s, the success probability of $\hat{H}$ approaches 1 as the sample size n increases indefinitely, or in symbol:

$$\mathbb{P}_s\left(\hat{H}(\mathbf{X}_1, \ldots, \mathbf{X}_n) = H_s\right) \to 1 \quad \text{as} \quad n \to \infty\,.$$

If learning method $\hat{H}$ converges to the truth in every state in the given state space $\mathcal{S}$, say that it is **statistically consistent**, or more intuitively, say that it converges to the truth **everywhere** with respect to the given learning problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$. Statistical consistency alone may not be enough for making a good learning method, but it is usually taken as one of the *minimal qualifications* for making a good learning method if this qualification can possibly be met. Indeed, there is a familiar, higher evaluation standard: A learning method $\hat{H}$ for tackling a causal learning problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ is said to converge to the truth with **global uniformity** (aka **uniform consistency**) if

$$\inf_{s \in \mathcal{S}} \mathbb{P}_s\left(\hat{H}(\mathbf{X}_1, \ldots, \mathbf{X}_n) = H_s\right) \to 1 \quad \text{as} \quad n \to \infty\,.$$

When the state space $\mathcal{S}$ is too inclusive, everywhere convergence can be easily unachievable, let alone the stronger condition of globally uniform convergence. This can already be seen from the example illustrated in figure 2. To be more specific, consider the joint distribution P^* parameterized

by $(p_1, p_2, p_3) = \left(\frac{1}{2}, \frac{3}{4}, \frac{7}{8}\right)$, as indicated in that figure. Suppose that the state space $\mathcal{S}$ contains at least the unfaithful causal state $s = (G_{\text{left}}, P^*)$ on the left and its counterpart $s' = (G_{\text{right}}, P^*)$ on the right; they share the same joint distribution P^*. Same joint distribution, same sampling distribution; so $\mathbb{P}_s = \mathbb{P}_{s'}$. Now, suppose that some hypothesis in $\mathcal{H}$ is true in one of those two states but false in the other. Then any learning method for tackling the present problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ has to fail to converge to the truth in one of those two causal states. That is, the problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ just described is too hard to allow the possibility of everywhere convergence, let alone the stronger mode of convergence, globally uniform convergence.

When high standards are unachievable, it is natural to look for lower standards and see whether they are achievable. So, define some weaker modes of convergence as follows.

The domain of convergence should be extended as far as possible. Accordingly, a learning method $\hat{H}$ for tackling a problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ is said to converge (in probability) to the truth **on a maximal domain** if no other learning method converges (in probability) to the truth in at least all causal states in $\mathcal{S}$ where $\hat{H}$ does and in strictly more causal states in $\mathcal{S}$.

The domain of convergence should, if possible, be extended to cover at least "almost everywhere", which will be defined quite standardly as in geometry and topology: "almost everywhere" as "everywhere" except on a "nowhere dense" subspace. Consider a very standard metric for measuring the distance between probability measures, the **total variation distance**, which is defined by: $\Delta(P, P') = \sup_A |P(A) - P'(A)|$, for any probability measures P and P'. Choose a metric δ defined on the set of all causal graphs over $\mathcal{V}$ (any metric would do), hold δ fixed, and let $\delta(G, G')$ measure the distance between two causal graphs G and G'. The distance between two causal states $s = (G, P)$ and $s' = (G', P')$ will be measured by a certain metric d, defined by $d(s, s') = \delta(G, G') + \Delta(P, P')$, i.e., the sum of the distance between the two causal structures and the distance between the two joint distributions. An **open ball** centered at a causal state s is a set taking this form:

$$B_\epsilon(s) \quad = \quad \left\{ s' \in \mathcal{S} : d(s, s') < \epsilon \right\} ,$$

where the radius ϵ is required to be positive. This turns the state space $\mathcal{S}$ into a topological space, whose **open sets** are defined as unions of open balls. The specific distance functions used to define the open sets are inessential to this paper; it is the open sets that are essential.[1] With respect to a topological space held fixed, a subset X is said to be **(topologically) negligible**, aka **nowhere dense**, if for every open set/ball B, there is some open set/ball B' that is nested within B and disjoint from X. In that case, X has an open "hole" B' in every local neighborhood B in the topological space; it is like a slice of Swiss cheese incredibly full of open holes. A learning method $\hat{H}$ for tackling a problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ is said to converge (in probability) to the truth **almost everywhere** if $\hat{H}$ converges (in probability) to the truth in all causal states in $\mathcal{S}$ except on a nowhere dense subset of $\mathcal{S}$.

Now we turn to uniformity of convergence. A learning method $\hat{H}$ for tackling a problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ is said to converge (in probability) to the truth with **adherent local uniformity** if, for any causal state $s \in \mathcal{S}$, if $\hat{H}$ converges (in probability) to the truth in s, then $\hat{H}$ converges (in probability) to the truth uniformly on some open neighborhood $B_\epsilon(s)$ of s in the state space $\mathcal{S}$, or in

1. We are indebted to Kevin T. Kelly for suggesting to us this topology, which significantly improves on the topology used in an earlier draft of this paper.

symbol, there exists a radius $\epsilon > 0$ such that

$$\inf_{s' \in B_\epsilon(s)} \mathbb{P}_{s'}\left(\hat{H}(\mathbf{X}_1, \ldots, \mathbf{X}_n) = H_{s'}\right) \to 1 \quad \text{as} \quad n \to \infty\,.$$

This means that, in every causal state in which the learning method converges to the truth, the error probability can be made not just low but *stably* low: the performance of such a learning method remains good *even when* the true, unknown CBN is vulnerable to a sufficiently small perturbation over which we do not have control. As we will see below from lemma 3, when a perturbation is sufficiently small, it will be a perturbation that changes the joint distribution only slightly but without a change in the causal structure—that is, with *adherence* to the unchanged causal structure.

4. Main Result & Discussion

Our main result addresses the question of where the domain of convergence should be extended if it cannot be extended to cover everywhere. Accordingly, when a learning method converges to the truth in state s, say that it has the convergence property be **secured** in s; otherwise say that it has the convergence property be **sacrificed** in s. Following Spirtes et al. (2000), we will focus on the task of learning a specific kind of causal hypothesis, called Markov equivalence hypothesis, which can be defined as follows. Two graphs G and G' are said to be **Markov equivalent** if $\mathcal{I}(G) = \mathcal{I}(G')$—that is, if G and G' are graphically so similar that they entail exactly the same conditional independence statements. For example, the two graphs G_{left} and G_{right} depicted in figure 2 are *not* Markov equivalent, for the right one entails the independence between X and Z, which is not entailed by the left one. Each graph G generates a **Markov equivalence class** $[G]$, defined as the set of all the graphs that are Markov equivalent to G. Each such class $[G]$ generates a **Markov equivalence hypothesis** H_G: "The causal graph of the true CBN is in Markov equivalence class $[G]$." The rationale for focusing on this kind of hypothesis will be explained below (in the discussion that follows the statement of the main result).

Theorem 1 *Let $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ be any causal learning problem such that $\mathcal{V}$ is a finite set of categorical variables, $\mathcal{S}$ is the state space consisting of all causal states on $\mathcal{V}$ (i.e., all causal Bayesian networks on $\mathcal{V}$), and $\mathcal{H}$ is the hypothesis set consisting of all the Markov equivalence hypotheses about $\mathcal{V}$. Suppose that there are at least two variables in $\mathcal{V}$. Then we have:*

1. *Learning problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ is so hard that it admits of no learning method that achieves the standard of everywhere convergence to the truth (as Spirtes et al. 2000 have already shown), let alone the higher standard of global uniform convergence.*

2. *But learning problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ is at least easy enough to admit of a learning method that achieves this lower standard: convergence to the truth (a) almost everywhere, (b) on a maximal domain, and (c) with adherent local uniformity.*

3. *For any learning method tackling problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$, if it achieves at least that much, namely the joint mode of convergence (a)+(b)+(c), then it has the convergence property be*

   3.1 *secured in (at least) every faithful causal state in $\mathcal{S}$,*

   3.2 *sacrificed in (at least) every unfaithful causal state in $\mathcal{S}$ that shares its joint probability distribution with some faithful causal state.*

The impossibility result in clause 1 is familiar, as mentioned above.[2] The possibility result in clause 2 is proved by a (long) sequence of constructions and verifications, detailed in appendix A. Clause 3 follows immediately from theorem 2, to be presented in the next section and proved below (in section 6 and appendix B). It is also possible to draw some pictures to illustrate why clause 3 holds in certain special cases; see appendix C for details.

The three clauses of this theorem are best understood with the help of figure 1: every joint mode of convergence outside the shaded area is unachievable, thanks to clause 1; every one inside the shaded area is achievable, thanks to clause 2. So the first two clauses determine the highest achievable mode of convergence considered in this paper—the one marked with a star in figure 1. Then, clause 3 says what it takes to achieve the highest achievable one. The rest of this section takes a closer look at the three clauses of this theorem in turn.

Clause 1 reports the impossibility of securing the convergence property everywhere, so sacrifices have to be made *somewhere*. The question is: *Where should sacrifices be made?* The crux of the matter is that the learning problem in question is very hard, so much so that (almost) all familiar standards for evaluating learning methods are too high an ideal to be achievable. In that case, there appears to be no achievable evaluation standard on the table, and hence no constraint on the candidate pool of good learning methods—in the present case, there appears to be no constraint on where the convergence property should be sacrificed. This is the source of the problem. In response, this paper pursues a general strategy: When confronted with a learning problem that is too hard to make it possible to achieve any of the familiar, higher evaluation standard (clause 1 of theorem 1), we should first proceed by looking for a lower evaluation standard that is desirable and achievable (clause 2 of theorem 1); if we manage to find one, we should then determine what it takes for a learning method to achieve that lower standard (clause 3 of theorem 1).

So, underlying clause 2 is the task of seeking a lower, achievable standard for the evaluation of causal learning methods. When everywhere convergence is impossible, it is still desirable to extend the domain of convergence as far as possible, preferably missing only a region that is negligible in a mathematically rigorous sense. This desideratum is made precise in terms of the modes of convergence (a) and (b) mentioned in clause 2. When globally uniform convergence is impossible—that is, when a high success probability cannot be obtained and retained under any perturbation, it is still desirable, if possible, to retain it under any perturbation of a limited kind. This desideratum is made precise in terms of the mode of convergence (c) mentioned in clause 2. It turns out that those weaker modes of convergence, (a), (b), and (c), are not just each achievable, but jointly achievable, as clause 2 shows.

To clarify, when it is possible to simultaneously achieve the three proposed modes of convergence, this achievement is only necessary, rather than sufficient, for making a good learning method. So, if there is any additional desideratum that can be achieved jointly with those three, it should be added to the stock of the evaluation standards in use—in order to further constrain the candidate pool for good learning methods. The point is that, as clause 3 shows, those three modes of convergence already work together to impose an interesting, significant constraint: to achieve *at least* those three simultaneously, a learning method based on non-experimental data has to secure the convergence property in at least every faithful causal state—this is mandatory rather than optional. So clause 3 justifies the standard design practice of sacrificing the convergence property *only* in unfaithful causal states.

2. This is the only clause whose truth depends on the assumption that there are at least two variables in $\mathcal{V}$.

The above illustrates how clauses 2 and 3 work together to address the question left by clause 1, the question of where the convergence property should be secured or sacrificed. The more traditional reaction to the impossibility result in clause 1 is to make the assumption that the true, unknown CBN is faithful. This restricts the state space $\mathcal{S}$ to the set $\mathcal{S}_{\text{faith}}$ of all faithful causal states, which, in a sense, restores the possibility of statistical consistency: everywhere convergence is achievable with respect to the modified, easier learning problem $(\mathcal{V}, \mathcal{S}_{\text{faith}}, \mathcal{H})$, as Spirtes et al. (2000) show. But to simply assume that the true, unknown CBN is faithful is to take for granted a specific answer to the question of where sacrifices should be made. The proposal of this paper is that causal learning theory need not be developed on the assumption that the true, unknown CBN is faithful, or on any other variant of the faithfulness assumption. Instead, the problem posed by clause 1 can, and should, be addressed by the general strategy that underlies clauses 2 and 3: look for what can be achieved, and achieve the best we can have. So, although the above theorem is stated in an unusual way, with the first clause being nothing but a familiar result, it is so stated in order to emphasize that the difficulty posed by clause 1 should be addressed by something like clauses 2 and 3.

The above theorem is limited in some ways. First, it only concerns a specific kind of hypothesis space: the set of the *Markov equivalence* hypotheses about the given set $\mathcal{V}$ of variables. This choice of a hypothesis space is made in this paper for a reason. A causal learning problem can certainly have a hypothesis space of some other kind. For example, Shimizu, Hoyer, Hyvärinen, and Kerminen (2006) study the causal learning problem $(\mathcal{V}', \mathcal{S}', \mathcal{H}')$ such that $\mathcal{V}'$ is a set of continuous variables and the state space $\mathcal{S}'$ is restricted to the (so-called) linear non-Gaussian acyclic causal models (or CBNs in which each variable is required to be a linear function of its parents plus an random noise term whose distribution must be non-Gaussian, but see Zhang and Hyvärinen (2009) for the extent to which this restriction might be relaxed). Under such parametric assumptions, it is unnecessary to restrict attention to Markov equivalence hypotheses or care much about the faithfulness condition. By way of contrast, it is the need to learn at least the true Markov equivalence hypothesis without making strong parametric assumptions that motivates Spirtes et al. (2000) to make the faithfulness assumption—to assume away unfaithful CBNs from the state space under consideration. So, to explore the possibility of developing causal learning theory without assuming away any unfaithful CBNs, a good starting point is to study the task of learning the true Markov equivalence hypothesis, as pursued in this paper.

The above theorem is also limited in another way: clause 3 only says that the convergence property has to be secured in *at least* all the faithful causal states. This raises some questions: *Exactly* where does the convergence property have to be secured? Also, *exactly* where does it have to be sacrificed? And *exactly* where is the sacrifice only optional but not mandatory? These questions are answered by a strengthening of the above theorem, to be presented in the next section.

5. Main Result Strengthened

To strengthen clause 3 of the preceding theorem, some more definitions are required.

A condition weaker than faithfulness is called (Pearl's) minimality (Pearl, 2009). Say that G is **minimal** to P if there exists no graph G' such that $\mathcal{I}(G) \subset \mathcal{I}(G') \subseteq \mathcal{I}(P)$. Call a causal state (G, P) **minimal** if G is minimal to P. The term 'minimal' can be understood intuitively this way. Suppose that P is the true joint probability distribution. So the conditional independence *facts* are those in $\mathcal{I}(P)$. Let's try to explain (some of) those facts by postulating a causal structure G. Assuming the causal Markov condition, we have to postulate a causal graph G with $\mathcal{I}(G) \subseteq \mathcal{I}(P)$.

So, of the conditional independence facts in $\mathcal{I}(P)$, those included in $\mathcal{I}(G)$ are explained (namely, entailed) by the postulated causal structure G but those in $\mathcal{I}(P) \setminus \mathcal{I}(G)$ are left unexplained (yet)—at least they cannot be explained by (lack of) causation if we postulate G. If we postulate a causal structure G that is minimal to P, it means that only a *minimal* set of conditional independence facts is left unexplained by (lack of) causation. In the limiting case that $\mathcal{I}(G) = \mathcal{I}(P)$, no conditional independence fact is left unexplained.

Call a causal state (G, P) **u-minimal (unambiguously minimal)** if G is minimal to P and every graph minimal to P is Markov equivalent to G. As indicated by the Venn diagram in figure 3, faithfulness is strictly stronger than u-minimality, which is in turn strictly stronger than minimality (Zhang, 2013).[3]

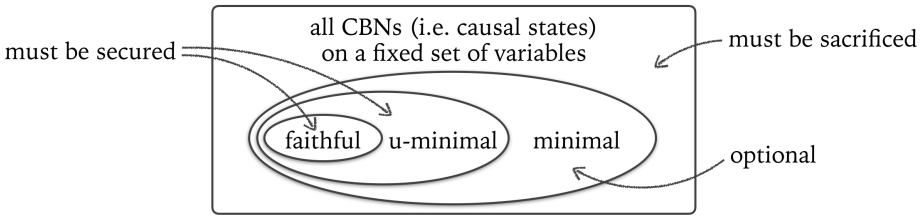


Figure 3: Summary of theorem 2

Theorem 2 *Continuing from theorem 1, we have: to achieve at least the joint mode of convergence* (a)+(b)+(c)*, the convergence property must be secured in all u-minimal causal states, must be sacrificed in all non-minimal causal states, and is optional for the other causal states (as summarized in figure 3). To be more precise:*

1. *For any learning method tackling problem* $(\mathcal{V}, \mathcal{S}, \mathcal{H})$*, if it achieves the joint mode of convergence* (a)+(b)+(c)*, then it has the convergence property be*
   - 1.1 *secured in all u-minimal causal states (including all faithful causal states) in* $\mathcal{S}$*,*
   - 1.2 *sacrificed in all causal states in* $\mathcal{S}$ *that are not minimal.*
2. *For any other causal state* s *in* $\mathcal{S}$ *(i.e. minimal but not u-minimal), some but not all learning methods achieving* (a)+(b)+(c) *converge to the truth in* s*.*

The proof is in appendix B.

But here is the strategy that underlies the proof. It is first shown that we only need to use the two modes of convergence (a) "almost everywhere" and (c) "adherently locally uniform" to force the convergence property to be sacrificed in all causal states that are not minimal. This means that (a)+(c) alone suffices to prove clause 1.2. This also means that, if the convergence property must be secured somewhere, (a)+(c) already requires that it be secured *only within* the range of the minimal causal states. Then, within that particular range, the domain of convergence is extended as much as possible in order to achieve the mode of convergence (b) "maximal domain". It is shown that, whenever the domain of convergence has not been extended enough to cover the set of all u-minimal

3. For an example of a u-minimal but not faithful causal state, see Zhang (2013, pp. 433–434). For examples of minimal but not u-minimal causal states, see Zhang (2013, p. 431).

causal states, it can always be extended *further* to do so. This means that, within the range of the minimal causal states, any maximal domain of convergence must cover at least the set of all u-minimal causal states, which leads to a proof of clause 1.1. To recap: modes (a) and (c) are used to prove clause 1.2, which is then used together with mode (b) to prove clause 1.1. Clause 2 makes a pair of existence claims for each causal state s that is minimal but not u-minimal: some of the learning methods that achieve (a)+(b)+(c) converge to the truth in s, but some others do not. Those two existence claims are proved with the help of the techniques developed for proving the existence claim (clause 2) of theorem 1. See appendix B for the complete proof.

6. Proof of an Important Lemma

This section states and proves what we call *the sacrifice lemma* (lemma 5 below), whose proof is particularly revealing because it *explains* why sacrificing the convergence property in certain causal states is a necessary cost of something good. This will lead to a proof of clause 1.2 of theorem 2.

Lemma 3 *Suppose that $\mathcal{V}$ is a finite set of variables, and that $\mathcal{S}$ is the set of all causal states (i.e., CBNs) on $\mathcal{V}$. With respect to the topological structure defined above (in section 3), we have: there exists a (sufficiently small) radius $\epsilon^* > 0$ such that, for any causal state $s = (G, P) \in \mathcal{S}$, the open ball $B_{\epsilon^*}(s)$ centered at s with radius ϵ^* contains only causal states that share the same causal graph, namely G.*

Proof Let δ be any metric chosen (in section 3) to measure the distances between causal graphs, and let $\delta_{\min}$ be the minimal distance measured by δ between two distinct causal graphs on $\mathcal{V}$. We have that $\delta_{\min} > 0$, for two reasons: first, δ as a metric must assign a nonzero distance to any pair of distinct causal graphs; second, there are only finitely many causal graphs on the finite set $\mathcal{V}$. Let the sought radius ϵ^* be $\delta_{\min}$. Consider any two causal states $s = (G, P)$ and $s' = (G', P')$ on $\mathcal{V}$ that are less ϵ^*-away from each other; that is, $d(s, s') < \epsilon^*$. It suffices to show that $G = G'$. Note that $\delta(G, G') + \Delta(P, P') = d(s, s') < \epsilon^* = \delta_{\min}$. Hence $\delta(G, G')$ is less than $\delta_{\min}$, the minimal distance between two distinct causal graphs on $\mathcal{V}$. So $G = G'$, as desired. ■

With respect to a causal learning problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$, the **domain of convergence** of a learning method is the set of the causal states in $\mathcal{S}$ in which that learning method converges (in probability) to the truth. Then we have:

Lemma 4 *If a learning method tackling a causal learning problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ converges to the truth almost everywhere, then its domain of convergence is dense in $\mathcal{S}$. If it converges to the truth with adherent local uniformity, then its domain of convergence is open in $\mathcal{S}$.*

Proof Immediate from definitions. ■

Having a dense and open domain of convergence implies a significant constraint on where the convergence property must be sacrificed:

Lemma 5 (The Sacrifice Lemma) *Let $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ be a causal learning problem such that $\mathcal{V}$ is a finite set of variables (whether or not those variables are categorical, discrete, or continuous), that $\mathcal{S}$ is the set of all causal states on $\mathcal{V}$, and that $\mathcal{H}$ is the set of the Markov equivalence hypotheses*

about $\mathcal{V}$. For any learning method $\hat{H}$ tackling problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$, if $\hat{H}$ has a dense and open domain of convergence on the state space $\mathcal{S}$, then $\hat{H}$ sacrifices the convergence property in every causal state in $\mathcal{S}$ that is not minimal.

Proof Let $\hat{H}$ be a causal learning method with a dense and open domain of convergence. By lemma 3, there exists a (small) radius $\epsilon^* > 0$ such that, if any two causal states are less than ϵ^*-away from each other, they share the same causal structure. Suppose, for *reductio*, that $\hat{H}$ converges to the truth in some non-minimal causal state $s_0 = (G, P) \in \mathcal{S}$. Since the domain of convergence is open (by hypothesis), there exists a nonzero radius $\epsilon \leq \epsilon^*$ such that $\hat{H}$ converges to the truth in every causal state in the open ball $B_\epsilon(s_0)$. Note that the Markov equivalence hypothesis H_G is true in s_0, and hence true in every causal state in the open ball $B_\epsilon(s_0)$, because $\epsilon \leq \epsilon^*$. (See the upper left part of figure 4 for a picture of the present situation, in which the so-called non-minimal plane represents the set of the non-minimal causal states in which H_G is true). From s_0 let's construct causal states s_1, s_2, and s_3 in the following three steps (as represented by the three arrows in figure 4):

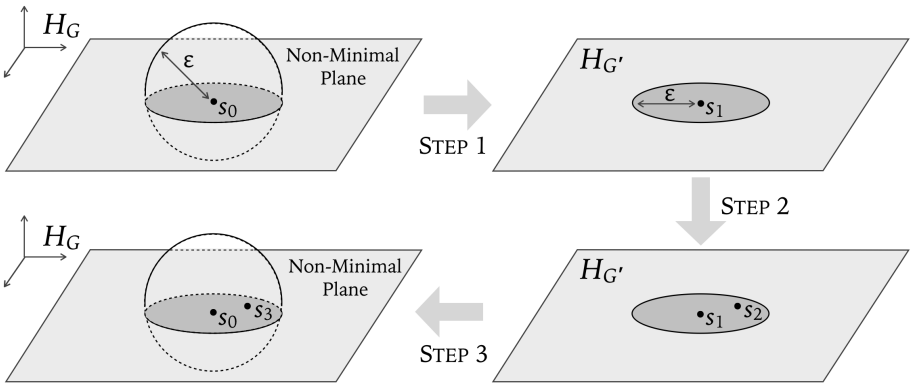


Figure 4: Constructions in the proof of lemma 5

- STEP 1. Since causal state $s_0 = (G, P)$ is not minimal, there exists a minimal causal state $s_1 = (G', P) \in \mathcal{S}$ with $\mathcal{I}(G) \subset \mathcal{I}(G') \subseteq \mathcal{I}(P)$. We have thus constructed s_1.
- STEP 2. Since $\mathcal{I}(G) \subset \mathcal{I}(G')$, G and G' are not Markov equivalent. So H_G and $H_{G'}$ are two distinct, incompatible hypotheses. Causal state $s_1 = (G', P)$ has an open neighborhood $B_\epsilon(s_1)$ with the same radius ϵ, in which all causal states share the same causal graph G' (because $\epsilon \leq \epsilon^*$). So $H_{G'}$ is true in every causal state in that open ball $B_\epsilon(s_1)$ (as depicted in the upper right part of figure 4). Then, since the domain of convergence is dense (by hypothesis), the open ball $B_\epsilon(s_1)$ contains at least one causal state in which $\hat{H}$ converges to the truth $H_{G'}$—now, choose one such causal state $s_2 = (G', P')$. We have thus constructed s_2.
- STEP 3. Take causal state $s_2 = (G', P')$, replace the graph therein by G to construct an ordered pair $s_3 = (G, P')$. Argue as follows that s_3 is indeed a causal state. Note that G is Markov to P' because $\mathcal{I}(G) \subseteq \mathcal{I}(G') \subseteq \mathcal{I}(P')$, where the first subset relation $\mathcal{I}(G) \subseteq \mathcal{I}(G')$ follows from the construction of G' and the second subset relation $\mathcal{I}(G') \subseteq \mathcal{I}(P')$ follows from the fact that $s_2 = (G', P')$ is a causal state (i.e., CBN), which must satisfy the Markov condition. Since G is Markov to P', $s_3 = (G, P')$ is indeed a causal state. We have thus constructed causal state s_3.

Causal state s_3 has some notable properties. First, s_3 is in the open ball $B_\epsilon(s_0)$, because $d(s_3, s_0) = \delta(G, G) + \Delta(P', P) = 0 + \Delta(P', P) = \delta(G', G') + \Delta(P', P) = d(s_2, s_1) < \epsilon$. Second, s_3 shares with s_2 the same joint distribution P' (and hence the same sampling distribution), so $\hat{H}$ converges to the same hypothesis in s_2 and in s_3, and that particular hypothesis is $H_{G'}$ (by the construction of s_2). It follows that $\hat{H}$ converges to a falsehood $H_{G'}$ in $s_3 = (G, P')$, which is in $B_\epsilon(s_0)$. Therefore, $\hat{H}$ fails to converge to the truth in some causal state in $B_\epsilon(s_0)$—contradiction. ■

Clause 1.2 of theorem 2 follows immediately from the previous two results: lemmas 4 and 5. Note that the above proof does not restrict the variables in $\mathcal{V}$ to be categorical variables. So, clause 1.2 of theorem 2 actually holds for any kinds of variables, be they categorical, discrete, or continuous.

To summarize, we submit that a causal learning method should, if possible, achieve at least the mode of convergence (a) "almost everywhere" plus the mode of convergence (c) "adherently locally uniform", which by lemma 4 implies having a dense and open domain of convergence, which by lemma 5 incurs a necessary cost: having the convergence property be sacrificed in every non-minimal causal state.

7. Closing: Some Possibilities for Future Research

The main results of this paper are theorems 1 and 2. Although they concern causal learning problems that involve only categorical variables, we conjecture that they can be generalized to cover some other causal learning problems, such as problems in which all causal states under consideration are linear Gaussian structural equation models. We also think that it should be possible to suitably generalize the main results to cover finite sets of variables of many different kinds. Our optimism is based on two observations. First, the key lemma 5 is applicable to any kind of variable. Second, with discrete or continuous variables that have an infinite range of possible values to take, the state space $\mathcal{S}$ can be too large to be captured by a finite-dimensional Euclidean space of parameters. In that case, it makes no sense to talk about mathematical negligibility as Lebesgue measure zero, which is popularized in the causal discovery community by Spirtes et al. (2000). But it still makes sense to understand mathematically negligible sets in topological terms, as nowhere dense sets or even meager sets (defined as unions of countably many nowhere dense sets). In fact, this is the main reason why we opt for the more applicable, topological conception of negligibility.

Acknowledgments

We are indebted to Kevin Kelly, Clark Glymour, Frederick Eberhardt, Christopher Hitchcock, Peter Spirtes, Kun Zhang, Konstantin Genin, and three anonymous referees for their very helpful comments on earlier drafts of this paper. Lin's research was supported by the University of California at Davis Startup Funds. Zhang's research was supported in part by the Research Grants Council of Hong Kong under the General Research Fund LU13600715, and by a Faculty Research Grant from Lingnan University.

References

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

William Feller. An introduction to probability theory and its applications. 1957.

Chris Meek. Strong-completeness and faithfulness in bayesian networks. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence. in:, Montreal, QU, Morgan Kaufmann, San Mateo, CA*, pages 411–418, 1995.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of machine learning research*, 7(Oct):2003–2030, 2006.

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

Zhalama, Jiji Zhang, and Wolfgang Mayer. Weakening faithfulness: some heuristic causal discovery algorithms. *International journal of data science and analytics*, 3(2):93–104, 2017.

Jiji Zhang. A comparison of three occam's razors for markovian causal models. *The British journal for the philosophy of science*, 64(2):423–448, 2013.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI press, 2009.

Appendix A. Proof of the Existence Result (Clause 2) of Theorem 1

The existence result (clause 2) of theorem 1 requires a long proof, which is broken down into three parts: We start with some topological preliminaries (appendix A.1), followed by some statistical preliminaries (appendix A.2), before we finally construct a learning method that witnesses the existence claim (appendix A.3).

Throughout this appendix, $\mathcal{V}$ is assumed to be a finite set of categorical variables, and $\mathcal{S}$ is the set of all causal states on $\mathcal{V}$.

A.1. Topological Preliminaries

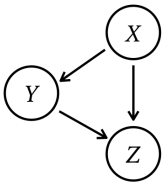
Let k denotes the number of assignments of values to all variables in $\mathcal{V}$. So, any joint distribution P of $\mathcal{V}$ is determined by the (joint) probabilities $p_1, p_2, \ldots, p_k$ that P distributes to those k assignments of values, respectively, and hence P can be identified with a point $(p_1, p_2, \ldots, p_k)$ in the k-dimensional Euclidean space $\mathbb{R}^k$. Note that a joint distribution P satisfies a conditional independence statement "$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$" if and only if the following equation holds:

$$\begin{aligned} & P(\mathbf{U}=\mathbf{u}, \mathbf{V}=\mathbf{v}, \mathbf{W}=\mathbf{w}) \cdot P(\mathbf{W}=\mathbf{w}) \\ & - P(\mathbf{U}=\mathbf{u}, \mathbf{W}=\mathbf{w}) \cdot P(\mathbf{V}=\mathbf{v}, \mathbf{W}=\mathbf{w}) \quad = \quad 0\,. \end{aligned}$$

Every term $P(\cdots)$ on the left side is a marginal probability, which can be expressed by a sum of some of the joint probabilities $p_1, p_2, \ldots, p_k$. So the left side can be expressed as a (second-degree) polynomial in k variables $p_1, p_2, \ldots, p_k$. More generally, every conditional independence

statement σ = "$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$" can be represented by a polynomial function $f_\sigma(x_1, x_2, \ldots, x_k)$ in k real-valued variables in this sense: a joint distribution P satisfies conditional independence statement σ = "$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$" if and only if $f_\sigma(p_1, p_2, \ldots, p_k) = 0$.

Hold a causal graph G fixed. Consider an arbitrary joint distribution P that satisfies the Markov condition with G (i.e., can form a CBN with G). It is well known that the joint distribution P factors according to the graph G into some conditional, marginal distributions. To be more specific, each joint probability p_i in P can be expressed as the product of some conditional probabilities, each of which is the probability for a variable to take a certain value conditional on its parents (in G) taking certain values (see the equations in figure 5 for an example with three binary variables). For



$P(X=1)$	$P(X=0)$
q_1	$1-q_1$

x	$P(Y=1 \mid X=x)$	$P(Y=0 \mid X=x)$
1	q_2	$1-q_2$
0	q_3	$1-q_3$

x	y	$P(Z=1 \mid X=x, Y=y)$	$P(Z=0 \mid X=x, Y=y)$
1	1	q_4	$1-q_4$
1	0	q_5	$1-q_5$
0	1	q_6	$1-q_6$
0	0	q_7	$1-q_7$

Factorization of a joint probability:

$P(X = 1, Y = 1, Z = 0)$

$= P(Z = 0 \mid Y = 1, X = 1) \cdot P(Y = 1 \mid X = 1) \cdot P(X = 1)$

$= (1 - q_4) \cdot q_2 \cdot q_1$

$0 \le q_1, q_2, \ldots, q_7 \le 1$

Figure 5: Conditional probability tables

convenience, we adopt the convention that, when a variable has no parent, its probability conditional on "its parents" means its unconditional probability. The conditional probabilities just mentioned are the real numbers in the conditional probability tables that are standardly used to represent a Bayesian network (see the tables in figure 5 for an example). So P is determined by its joint probabilities $p_1, p_2, \ldots, p_k$, each of which can then be expressed as a polynomial in some of the conditional probabilities $q_1, q_2, \ldots, q_m$, where each conditional probability q_i can take any value in the unit interval $[0, 1]$ and m is in general less than k (in figure 5, $m = 7 < k = 2^3 = 8$). So a conditional independence statement σ = "$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$" can be represented by another polynomial g_σ, so that

$$\begin{aligned} & P \text{ satisfies } \sigma = \text{"}\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}\text{"} \\ \Leftrightarrow \quad & f_\sigma(p_1, p_2, \ldots, p_k) = 0 \\ \Leftrightarrow \quad & g_\sigma(q_1, q_2, \ldots, q_m) = 0 \end{aligned}$$

More generally, let G be a causal graph on a finite set of categorical variables, and let $\mathcal{D}_G$ be the set of the joint distributions that are Markov to G (i.e., the joint distributions which can form a CBN with G). It is well known that the above provides a smooth parametrization of $\mathcal{D}_G$ by an

m-dimensional parameter space, the m-dimensional unit cube $[0,1]^m$; under this parametrization, every conditional independence statement is represented by a polynomial (Meek, 1995).

Lemma 6 *Let $\sigma =$ "$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$" be a conditional independence statement about $\mathcal{V}$. The joint distributions in $\mathcal{D}_G$ that violate $\sigma =$ "$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$" form an open subset of $\mathcal{D}_G$.*

Proof The joint distributions in $\mathcal{D}_G$ that violate σ form the set represented by $g_\sigma^{-1}[\mathbb{R} \setminus \{0\}]$, which is an open set because it is a pre-image of an open set ($\mathbb{R} \setminus \{0\}$) under a continuous function (the polynomial function g_σ). ■

Lemma 7 *Continuing from the preceding lemma, suppose further that G does not entail the conditional independence statement $\sigma =$ "$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$". Then the joint distributions in $\mathcal{D}_G$ that violate $\sigma =$ "$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$" form an open, dense subset of $\mathcal{D}_G$.*

Proof By the preceding lemma, the set of the joint distributions in $\mathcal{D}_G$ that violate $\sigma =$ "$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$" is open in $\mathcal{D}_G$. To show that this set is dense in $\mathcal{D}_G$, suppose for *reductio* that it is not dense. Then $\mathcal{D}_G$ has an open subset on which $\sigma =$ "$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$" is satisfied. Using the conditional probability parametrization described above, it follows that the m-dimensional unit cube $[0,1]^m$ has an open subset O on which $g_\sigma(q_1, q_2, \ldots, q_m) = 0$. It follows that g_σ is identically zero on O and all partial derivatives of g_σ are also identically zero on O, which implies that the Taylor series expansion of g_σ only has zero coefficients, which implies that g_σ is identically zero on the entire cube $[0,1]^m$. So the conditional independence statement $\sigma =$ "$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$ is satisfied by *all* distributions in $\mathcal{D}_G$. It follows that G entails $\sigma =$ "$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$—contradiction. ■

Lemma 8 *Let G be a causal graph on $\mathcal{V}$, and Σ be a finite set of (some, possibly not all) conditional independence statements that G does not entail. Then we have:*

1. *The joint distributions in $\mathcal{D}_G$ that violate every conditional independence statement in Σ form an open, dense subset of $\mathcal{D}_G$.*
2. *The joint distributions in $\mathcal{D}_G$ that satisfy at least one conditional independence statement in Σ form a nowhere dense subset of $\mathcal{D}_G$.*

Proof Clause 1 follows immediately from the preceding lemma and the following, familiar fact in general topology: open, dense subsets are closed under finite conjunctions. Clause 2 follows from clause 1 for two reasons: first, the set mentioned in clause 2 is the complement (in $\mathcal{D}_G$) of the set mentioned clause 1; second, it is a familiar fact in general topology that any complement of an open, dense, subset is a nowhere dense subset. ■

The above results concern spaces of joint distributions, and can be carried over to spaces of causal states as follows. Let $\mathcal{S}_G$ be the topological space of the causal states whose graphs are identical to G. A causal state $s = (G, P)$ is said to **satisfy** (or **violate**) a conditional independence statement σ if the underlying joint distribution P satisfies (or violates) σ.

Lemma 9 *Let G be a causal graph on $\mathcal{V}$, and Σ be a finite set of (some, possibly not all) conditional independence statements that G does not entail. Then we have:*

1. *The causal states in $\mathcal{S}_G$ that violate every conditional independence statement in Σ form an open, dense subset of $\mathcal{S}_G$.*

2. *The causal states in $\mathcal{S}_G$ that satisfy at least one conditional independence statement in Σ form a nowhere dense subset of $\mathcal{S}_G$.*

Proof Immediate from the previous lemma and the fact that $\mathcal{D}_G$ is homeomorphic to $\mathcal{S}_G$, with the homeomorphism: $P \mapsto (G, P)$, which maps each joint distribution P in $\mathcal{D}_G$ to a causal state (G, P) in $\mathcal{S}_G$. ■

Lemma 10 *For any causal graph G on $\mathcal{V}$, $\mathcal{S}_G$ is open in $\mathcal{S}$.*

Proof Immediate from lemma 3 (in section 6). ■

Lemma 11 *Let G be a causal graph on $\mathcal{V}$, and Σ be a set of (some, possibly not all) conditional independence statements that G does not entail. Then we have:*

1. *The causal states in $\mathcal{S}_G$ that violate every conditional independence statement in Σ form an open subset of $\mathcal{S}$.*

2. *The causal states in $\mathcal{S}_G$ that satisfy at least one conditional independence statement in Σ form a nowhere dense subset of $\mathcal{S}$.*

Proof Immediate from the previous two lemmas, together with the following, familiar facts in general topology: If a set is an open subset of an open subset of a space, it is open in the space. If a set is a nowhere dense subset of a subset of a space, it is nowhere dense in the space. ■

Proposition 12 *Every causal state $s = (G, P)$ in $\mathcal{S}$ has a (sufficiently small) open neighborhood such that, for any causal state $s' = (G', P')$ in that open neighborhood, $G' = G$ and $\mathcal{I}(P') \subseteq \mathcal{I}(P)$. Or in words, every causal state s in $\mathcal{S}$ has a (sufficiently small) open neighborhood in which every causal state shares with s the same causal graph and violates at least all the conditional independence statements that s violates.*

Proof Consider an arbitrary causal state $s = (G, P)$ in $\mathcal{S}$. Since $s = (G, P)$ is a causal state, we have that G is Markov to P, and it follows that G does not entail any conditional independence statement that P violates. This allows us to apply lemma 11 to graph G together with Σ being the set of the conditional independence statements that P violates—namely, those that s violates. Then, by clause 1 of lemma 11, we have: the set of the causal states in $\mathcal{S}_G$ that violate at least all the conditional independence statements that s violates is an open set in $\mathcal{S}$. This open set is an open neighborhood of $s = (G, P)$ with the sought properties. The present proposition follows. ■

Proposition 13 *In the space $\mathcal{S}$ of all causal states on $\mathcal{V}$, the set of the unfaithful ones is nowhere dense and so is the set of the non-minimal ones.*

Proof Applying the second clause of lemma 11 to any causal graph G on $\mathcal{V}$ together with Σ being the set of all conditional independence statements that G does not entail, we have: the unfaithful causal states in $\mathcal{S}_G$ form a set $\mathcal{S}_G^{\text{unf}}$, which is nowhere dense in $\mathcal{S}$. Then, the set of the unfaithful causal states in $\mathcal{S}$ is nowhere dense in $\mathcal{S}$, for two reasons: first, this set is the finite union of the nowhere dense subsets $\mathcal{S}_G^{\text{unf}}$ such that G is a causal graph on $\mathcal{V}$; second, nowhere dense subsets are closed under finite unions. Moreover, the set of the non-minimal causal states in $\mathcal{S}$ is also nowhere dense in $\mathcal{S}$, for two reasons: first, it is a subset of a nowhere dense set, namely, the set of the unfaithful ones; second, any subset of a nowhere dense set is nowhere dense. ■

A.2. Statistical Preliminaries

Lemma 14 (Hoeffding's Inequality for Empirical Measures) *Let $\hat{P}_n$ be the empirical distribution (namely, frequency counts) of n observations obtained by IID sampling from a categorical distribution P. Then, for any $\epsilon > 0$ and for any sample size n, we have:*

$$\mathbb{P}\left(\Delta(\hat{P}_n, P) < \epsilon\right) \quad \geq \quad 1 - 2^k e^{-2n\epsilon^2},$$

where $\mathbb{P}_s$ denotes the sampling distribution generated by P, namely the ∞-fold probability measure generated by P under the IID assumption, and Δ is the total variation distance, and k is a constant denoting the number of the categories of P.

Proof It is routine to prove this result in probability theory. Here is one of the standard forms of Hoeffding's Inequality:

$$\mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[\overline{X}_n]\right| \geq \epsilon\right) \quad \leq \quad 2e^{-2n\epsilon^2}.$$

Let $\mathcal{X}$ be the set of the k given categories. So the set of the relevant events is $2^{\mathcal{X}}$. Let $\overline{2^{\mathcal{X}}}$ be a subset of $2^{\mathcal{X}}$ constructed as follows: for every pair (A, A') of sets that form a partition of $\mathcal{X}$, choose exactly one of the two sets, A or A', and put it in $\overline{2^{\mathcal{X}}}$. Note that the cardinality of $\overline{2^{\mathcal{X}}}$ is 2^{k-1}. For each proposition $A \in 2^{\mathcal{X}}$, apply Hoeffding's inequality to $|\hat{P}_n(A) - \mathbb{E}[\hat{P}_n(A)]|$, which is equal to $|\hat{P}_n(A) - P(A)|$, so we have:

$$\mathbb{P}\left(\left|\hat{P}_n(A) - P(A)\right| \geq \epsilon\right) \quad \leq \quad 2e^{-2n\epsilon^2}.$$

Then we have:

$$\begin{aligned}
\mathbb{P}\left(\Delta(\hat{P}_n, P) \geq \epsilon\right) &= \mathbb{P}\left(\max_{A \in 2^{\mathcal{X}}} \left|\hat{P}_n(A) - P(A)\right| \geq \epsilon\right) \\
&= \mathbb{P}\left(\bigvee_{A \in 2^{\mathcal{X}}} \left(\left|\hat{P}_n(A) - P(A)\right| \geq \epsilon\right)\right) \\
&= \mathbb{P}\left(\bigvee_{A \in \overline{2^{\mathcal{X}}}} \left(\left|\hat{P}_n(A) - P(A)\right| \geq \epsilon\right)\right) \\
&\leq \sum_{A \in \overline{2^{\mathcal{X}}}} \mathbb{P}\left(\left|\hat{P}_n(A) - P(A)\right| \geq \epsilon\right) \\
&\leq \sum_{A \in \overline{2^{\mathcal{X}}}} 2e^{-2n\epsilon^2} \quad = \quad 2^{k-1} \cdot 2e^{-2n\epsilon^2} \quad = \quad 2^k e^{-2n\epsilon^2}.
\end{aligned}$$

So $\mathbb{P}\left(\Delta(\hat{P}_n, P) < \epsilon\right) \geq 1 - 2^k e^{-2n\epsilon^2}$, as required. ■

Proposition 15 *Every conditional independence statement $\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$ that involves only categorical variables has a test T with the following convergence properties:*

1. *on the space of all possible distributions of $\mathbf{X} = \mathbf{U} \cup \mathbf{V} \cup \mathbf{W}$ that satisfy the independence statement $\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$, test T converges to the truth everywhere with global uniformity;*
2. *on the space of all possible distributions of $\mathbf{X} = \mathbf{U} \cup \mathbf{V} \cup \mathbf{W}$ that violate the independence statement $\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$, test T converges to the truth everywhere with local uniformity.*

Proof Let $\mathbf{U}, \mathbf{V}, \mathbf{W}$ be three disjoint sets of categorical variables. We are going to test the hypothesis that $\mathbf{U}$ and $\mathbf{V}$ are independent given $\mathbf{W}$. Consider an arbitrary joint probability distribution P of $\mathbf{X} = \mathbf{U} \cup \mathbf{V} \cup \mathbf{W}$. Let $P(\mathbf{u}, \mathbf{v}, \mathbf{w})$ abbreviate $P(\mathbf{U} = \mathbf{u}, \mathbf{V} = \mathbf{v}, \mathbf{W} = \mathbf{w})$, and similarly for $P(\mathbf{u}, \mathbf{v})$, $P(\mathbf{w})$, $P(\mathbf{x})$ etc. Define the following L_1-distance of P from the independence of $\mathbf{U}$ and $\mathbf{V}$ given $\mathbf{W}$:

$$L_1(P) \quad = \quad \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \left| P(\mathbf{u}, \mathbf{v}, \mathbf{w})P(\mathbf{w}) - P(\mathbf{u}, \mathbf{w})P(\mathbf{v}, \mathbf{w}) \right|,$$

where $\mathbf{u}, \mathbf{v}, \mathbf{w}$ range over the possible values of $\mathbf{U}, \mathbf{V}, \mathbf{W}$, respectively. Let $\hat{P}_n$ be the empirical distribution (namely, frequency counts) of n observations. (So $\hat{P}_n$ is a random probability distribution of $\mathbf{X} = \mathbf{U} \cup \mathbf{V} \cup \mathbf{W}$.) It suffices to prove that the existence claim is witnessed by the following test:

- Accept the hypothesis of conditional independence if $L_1(\hat{P}_n) < \frac{1}{n^{1/4}}$.
- Reject that hypothesis otherwise.

We will need to bound $|L_1(P) - L_1(Q)|$, where P and Q are two arbitrary probability distributions of $\mathbf{X} = \mathbf{U} \cup \mathbf{V} \cup \mathbf{W}$. Bound it as follows:

$$\begin{aligned}
& |L_1(P) - L_1(Q)| \\
= \; & \Big| \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \left| P(\mathbf{u}, \mathbf{v}, \mathbf{w})P(\mathbf{w}) - P(\mathbf{u}, \mathbf{w})P(\mathbf{v}, \mathbf{w}) \right| - \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \left| Q(\mathbf{u}, \mathbf{v}, \mathbf{w})Q(\mathbf{w}) - Q(\mathbf{u}, \mathbf{w})Q(\mathbf{v}, \mathbf{w}) \right| \Big| \\
\leq \; & \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \Big| \left| P(\mathbf{u}, \mathbf{v}, \mathbf{w})P(\mathbf{w}) - P(\mathbf{u}, \mathbf{w})P(\mathbf{v}, \mathbf{w}) \right| - \left| Q(\mathbf{u}, \mathbf{v}, \mathbf{w})Q(\mathbf{w}) - Q(\mathbf{u}, \mathbf{w})Q(\mathbf{v}, \mathbf{w}) \right| \Big| \\
\leq \; & \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \Big(\left| P(\mathbf{u}, \mathbf{v}, \mathbf{w})P(\mathbf{w}) - Q(\mathbf{u}, \mathbf{v}, \mathbf{w})Q(\mathbf{w}) \right| + \left| P(\mathbf{u}, \mathbf{w})P(\mathbf{v}, \mathbf{w}) - Q(\mathbf{u}, \mathbf{w})Q(\mathbf{v}, \mathbf{w}) \right| \Big) \\
& \text{by} \big| |a - b| - |a' - b'| \big| \leq |a - a'| + |b + b'| \\
= \; & \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \left| P(\mathbf{u}, \mathbf{v}, \mathbf{w})P(\mathbf{w}) - Q(\mathbf{u}, \mathbf{v}, \mathbf{w})Q(\mathbf{w}) \right| + \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \left| P(\mathbf{u}, \mathbf{w})P(\mathbf{v}, \mathbf{w}) - Q(\mathbf{u}, \mathbf{w})Q(\mathbf{v}, \mathbf{w}) \right|.
\end{aligned}$$

Then we are going to bound the first and second terms, respectively. Bound the first term as follows:

$$\begin{aligned}
& \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \big|P(\mathbf{u},\mathbf{v},\mathbf{w})P(\mathbf{w}) - Q(\mathbf{u},\mathbf{v},\mathbf{w})Q(\mathbf{w})\big| \\
= & \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \Big|\big(P(\mathbf{u},\mathbf{v},\mathbf{w}) - Q(\mathbf{u},\mathbf{v},\mathbf{w})\big)\cdot P(\mathbf{w}) + Q(\mathbf{u},\mathbf{v},\mathbf{w})\cdot\big(P(\mathbf{w}) - Q(\mathbf{w})\big)\Big| \\
\leq & \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \Big|\big(P(\mathbf{u},\mathbf{v},\mathbf{w}) - Q(\mathbf{u},\mathbf{v},\mathbf{w})\big)\cdot P(\mathbf{w})\Big| + \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \Big|Q(\mathbf{u},\mathbf{v},\mathbf{w})\cdot\big(P(\mathbf{w}) - Q(\mathbf{w})\big)\Big| \\
\leq & \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \big|P(\mathbf{u},\mathbf{v},\mathbf{w}) - Q(\mathbf{u},\mathbf{v},\mathbf{w})\big| + \sum_{\mathbf{w}} \Big(\big|P(\mathbf{w}) - Q(\mathbf{w})\big|\cdot \sum_{\mathbf{u},\mathbf{v}} Q(\mathbf{u},\mathbf{v},\mathbf{w})\Big) \\
\leq & \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \big|P(\mathbf{u},\mathbf{v},\mathbf{w}) - Q(\mathbf{u},\mathbf{v},\mathbf{w})\big| + \sum_{\mathbf{w}} \Big(\big|P(\mathbf{w}) - Q(\mathbf{w})\big|\cdot Q(\mathbf{w})\Big) \\
\leq & \sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \big|P(\mathbf{u},\mathbf{v},\mathbf{w}) - Q(\mathbf{u},\mathbf{v},\mathbf{w})\big| + \sum_{\mathbf{w}} \big|P(\mathbf{w}) - Q(\mathbf{w})\big| \\
\leq & \sum_{\mathbf{x}} \big|P(\mathbf{x}) - Q(\mathbf{x})\big| + \sum_{\mathbf{x}} \big|P(\mathbf{x}) - Q(\mathbf{x})\big| \\
= & \; 2\cdot \sum_{\mathbf{x}} \big|P(\mathbf{x}) - Q(\mathbf{x})\big| \\
= & \; 4\,\Delta(P,Q)\,.
\end{aligned}$$

The last step follows because $\sum_{\mathbf{x}} \big|P(\mathbf{x}) - Q(\mathbf{x})\big| = 2\Delta(P,Q)$, which is a consequence of the fact that $\Delta(P,Q)$ denotes the total variation distance between P and Q. The second term can be bounded in the same way:

$$\sum_{\mathbf{u},\mathbf{v},\mathbf{w}} \big|P(\mathbf{u},\mathbf{w})P(\mathbf{v},\mathbf{w}) - Q(\mathbf{u},\mathbf{w})Q(\mathbf{v},\mathbf{w})\big| \;\leq\; 4\,\Delta(P,Q)\,.$$

So $|L_1(P) - L_1(Q)|$ can be bounded neatly as follows:

$$|L_1(P) - L_1(Q)| \;\leq\; 8\,\Delta(P,Q)\,. \tag{1}$$

Let P be the (unknown) true distribution under the null hypothesis that the conditional independence statement holds. So $L_1(P) = 0$. Let $\hat{P}_n$ be the random empirical distribution generated from P with sample size n. Consider the following inequality:

$$L_1(\hat{P}_n) \;=\; \big|L_1(\hat{P}_n) - L_1(P)\big| \;\leq\; 8\,\Delta(\hat{P}_n, P) \;<\; \frac{1}{n^{1/4}}\,.$$

This inequality holds with a probability at least $1 - 2^k e^{-2n\left(\frac{1}{8n^{1/4}}\right)^2} = 1 - 2^k e^{-\frac{\sqrt{n}}{32}}$ (by inequality (1) and lemma 14), which converges to 1 as n tends to infinity. Also note that this probability bound holds for all distributions under the null hypothesis. So clause 1 follows.

Now, let's turn to how the test performs under the alternative hypothesis that the conditional independence statement does not hold. Let P^* be the (unknown) true distribution under the alternative hypothesis. So $L_1(P^*) > 0$. Let P an arbitrary distribution in the open ball $B_{L_1(P^*)/32}(P^*)$.

Let $\hat{P}_n$ be the random empirical distribution generated from P with sample size n. Consider the following inequality:

$$\begin{aligned} L_1(\hat{P}_n) &\geq L_1(P^*) - \left|L_1(P^*) - L_1(P)\right| - \left|L_1(P) - L_1(\hat{P}_n)\right| \\ &\geq L_1(P^*) - 8\,\Delta(P^*, P) - 8\,\Delta(P, \hat{P}_n) \\ &> L_1(P^*) - 8\left(\frac{L_1(P^*)}{32}\right) - 8\left(\frac{L_1(P^*)}{16}\right) \\ &= \frac{1}{4}L_1(P^*) \\ &\geq \frac{1}{n^{1/4}} \end{aligned}$$

This inequality holds with a probability at least $1 - 2^k e^{-2n\left(\frac{L_1(P^*)}{16}\right)^2} = 1 - 2^k e^{-\frac{n}{128}L_1(P^*)^2}$, for any joint distribution P in the open ball $B_{L_1(P^*)/32}(P^*)$ and for any n large enough to guarantee that $\frac{1}{4}L_1(P^*) \geq \frac{1}{n^{1/4}}$ (by inequality (1) and lemma 14). Also note that this probability lower bound $1 - 2^k e^{-\frac{n}{128}L_1(P^*)^2}$ converges to 1 as n tends to infinity. So locally uniform convergence holds. This establishes clause 2. ■

The above proof actually establishes not just convergence in probability but also almost sure convergence, which follows from two things: the Borel-Cantelli lemma,[4] and the fact that the error probabilities in question converge to zero quickly enough so that they sum to a finite number. Indeed, under the null hypothesis, the sum of the error probabilities is $\sum_{n=1}^{\infty} 2^k e^{-\frac{\sqrt{n}}{32}} < \infty$. Under the alternative hypothesis, the sum of the error probabilities is $\sum_{n=1}^{\infty} 2^k e^{-\frac{n}{128}L_1(P^*)^2} < \infty$.

A.3. Construction of Learning Methods

Given a finite set $\mathcal{V}$ of variables, the following is an algorithm for constructing learning methods that will be shown to witness the existence claim in theorem 1.

Step 1. Let each conditional independence statement about $\mathcal{V}$ be associated with a test of it that achieves the convergence properties established in proposition 15. Combine those tests into a single "super" test T, which maps each data set $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ to the set $\Sigma = T(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of all the conditional independence statements accepted by their associated tests given data set $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

Step 2. Linearly order all Markov hypotheses about $\mathcal{V}$ into a sequence $H_{G_1}, H_{G_2}, \ldots, H_{G_k}$ such that $\mathcal{I}(G_i) \supset \mathcal{I}(G_j)$ implies $i < j$.

Step 3. Construct a function F that maps each set Σ of conditional independence statements about $\mathcal{V}$ to the first hypothesis H_{G_i} in the sequence such that $\mathcal{I}(G_i) \subseteq \Sigma$.

Step 4. Construct learning method $\hat{H} = F \circ T$.

A graph G on $\mathcal{V}$ is said to be **minimal** to a set Σ of conditional independence statements if there is no graph G' on $\mathcal{V}$ such that $\mathcal{I}(G) \subset \mathcal{I}(G') \subseteq \Sigma$.

4. For a review of Borel-Cantelli lemma, see chapter 14 of Feller (1957).

Lemma 16 *There is a learning method that can be constructed from the above procedure. Furthermore, any such learning method $\hat{H} = F \circ T$ has the following properties:*

1. *Whenever $F(\Sigma) = H_G$, then G is minimal to Σ.*
2. *Whenever $F(\Sigma) = H_G$, then $F(\Sigma') = H_G$ for any set Σ' with $\mathcal{I}(G) \subseteq \Sigma' \subseteq \Sigma$.*

Proof The existence of such a learning method follows from the following three facts. First, there exists a "super" test T of conditional independence with the property required in step 1 (by proposition 15). Second, there exists a sequence of causal hypotheses with the property required in step 2 (which is obvious because there are only finitely many hypotheses to be ordered). Finally, function F is well-defined (because, as an elementary result in the theory of Bayesian networks, for each set Σ of conditional independence statements about $\mathcal{V}$, there exists a graph G on $\mathcal{V}$ such that $\mathcal{I}(G) = \varnothing \subseteq \Sigma$).

Consider an arbitrary learning method $\hat{H}$ that can be constructed from the above procedure: $\hat{H} = F \circ T$, with a function F, a test T, and a sequence of causal hypotheses $H_{G_1}, H_{G_2}, \ldots, H_{G_k}$ satisfying all the required properties. Argue for the two clauses as follows.

To establish clause 1, suppose for *reductio* that $F(\Sigma) = H_G$ but G is not minimal to Σ, namely there is a graph G' on $\mathcal{V}$ such that $\mathcal{I}(G) \subset \mathcal{I}(G') \subseteq \Sigma$. Since the sequence $H_{G_1}, H_{G_2}, \ldots, H_{G_k}$ contains all the Markov equivalence hypotheses about $\mathcal{V}$, we have that $H_G = H_{G_j}$ and $H_{G'} = H_{G_i}$ for some $j, i \leq k$. So, to rewrite what we have already had: $F(\Sigma) = H_{G_j}$ and $\mathcal{I}(G_j) \subset \mathcal{I}(G_i) \subseteq \Sigma$. Since $\mathcal{I}(G_i) \supset \mathcal{I}(G_j)$, by the requirement in step 2 of the procedure we have that $i < j$. That is, H_{G_i} is a hypothesis that occurs earlier than H_{G_j} does in the sequence. But note that $\mathcal{I}(G_i) \subseteq \Sigma$. So, by the requirement in step 3, $F(\Sigma)$ is not H_{G_j} but must be either H_{G_i} or some earlier hypothesis in the sequence—contradiction. This establishes clause 1.

To establish clause 2, suppose that $F(\Sigma) = H_G$ and that $\mathcal{I}(G) \subseteq \Sigma' \subseteq \Sigma$. It suffices to show that $F(\Sigma') = H_G$. Since the sequence $H_{G_1}, H_{G_2}, \ldots, H_{G_k}$ contains all the Markov equivalence hypotheses about $\mathcal{V}$, we have that $F(\Sigma) = H_G = H_{G_i}$ and $\mathcal{I}(G) = \mathcal{I}(G_i)$ for some index i of the sequence. Since $F(\Sigma) = H_{G_i}$, by the requirement in step 3 we have:

(i) $\mathcal{I}(G_{i'}) \not\subseteq \Sigma$ for each $i' < i$.

Since $\mathcal{I}(G) = \mathcal{I}(G_i)$ and $\mathcal{I}(G) \subseteq \Sigma'$ (by hypothesis), we have:

(ii) $\mathcal{I}(G_i) \subseteq \Sigma'$.

Since (i) holds and $\Sigma' \subseteq \Sigma$ (by hypothesis), we have:

(iii) $\mathcal{I}(G_{i'}) \not\subseteq \Sigma'$ for each $i' < i$,

So, by (ii) and (iii) and the requirement in step 3, we have that $F(\Sigma') = H_{G_i}$. It follows that $F(\Sigma') = H_G$. This establishes clause 2. ■

Lemma 17 *For every u-minimal causal state $s = (G, P)$ and every learning method $\hat{H} = F \circ T$ that can be constructed from the above procedure, we have that $F(\mathcal{I}(P)) = H_G$.*

Proof Immediate from the requirements in steps 2 and 3. ■

The above lemma is the last one we need for proving clause 2 of theorem 1. The next lemma will be used to prove clause 3 of theorem 2.

Lemma 18 *For every minimal causal state $s = (G, P)$, there is a learning method $\hat{H} = F \circ T$ that can be constructed from the above procedure such that $F(\mathcal{I}(P)) = H_G$.*

Proof Let $s = (G, P)$ be any minimal causal state, and let $\mathcal{M}(P)$ denote the set of all Markov equivalence hypotheses whose graphs are minimal to $\mathcal{I}(P)$. Since s is minimal, we have: first, $H_G \in \mathcal{M}(P)$; second, for every $H_{G'} \in \mathcal{M}(P)$ distinct from H_G, $\mathcal{I}(G') \not\supset \mathcal{I}(G)$. Hence, there is a linear order of all the Markov equivalence hypotheses about $\mathcal{V}$, $H_{G_1}, H_{G_2}, \ldots, H_{G_k}$, such that (i) $\mathcal{I}(G_i) \supset \mathcal{I}(G_j)$ implies $i < j$, (ii) $H_G = H_{G_m}$ for some index m, and for every $H_{G'} \in \mathcal{M}(P)$ distinct from H_G, $H_{G'} = H_{G_n}$ for some index n and $m < n$. Thanks to (i), this linear order can be used in step 2 of the above procedure, which, by clause 1 of lemma 16, yields a learning method $\hat{H} = F \circ T$ such that $F(\mathcal{I}(P)) \in \mathcal{M}(P)$. Then, because of (ii) and the requirement of step 3 of the procedure, it follows that $F(\mathcal{I}(P)) = H_{G_m} = H_G$. ■

Proposition 19 *Let $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ be any causal learning problem such that $\mathcal{V}$ is a finite set of categorical variables, $\mathcal{S}$ is the state space consisting of all causal states on $\mathcal{V}$, and $\mathcal{H}$ is the hypothesis set consisting of all the Markov equivalence hypotheses about $\mathcal{V}$. Then there is a learning method for $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ that can be constructed from the above procedure, and any such learning method has the following properties:*

(a) *convergence to the truth almost everywhere,*

(b) *on a maximal domain,*

(c) *with adherent local uniformity.*

Proof Let $\hat{H}$ be a learning method that can be constructed from the above procedure: $\hat{H} = F \circ T$.

To prove property (a), note that $\hat{H}$ converges to the truth in every u-minimal state in $\mathcal{S}$, thanks to construction step 1, the convergence/consistency property of T established in proposition 15, and lemma 17. So $\hat{H}$ fails to converge to the truth *only* in states in $\mathcal{S}$ that are not u-minimal, but those states form a nowhere dense subset of $\mathcal{S}$ (thanks to proposition 13). So property (a) follows.

To prove property (b), consider an arbitrary learning method $\hat{H}'$ that converges to the truth in all states where $\hat{H}$ does. It suffices to show that $\hat{H}'$ does not converge to the truth in more states than $\hat{H}$ does. Let $(G, P) \in \mathcal{S}$ be a state in which $\hat{H}'$ converges to the truth. It suffices to show that $\hat{H}$ converges to the truth in (G, P). Recall that $\hat{H} = F \circ T$, and by construction step 3, that $F(\mathcal{I}(P)) = H_{G'}$ for some graph G' Markov to P. So (G', P) is a state in $\mathcal{S}$. Then, by proposition 15, $\hat{H}$ converges to the truth in state (G', P)—and, hence, $\hat{H}'$ does, too, by hypothesis. To sum up, $\hat{H}'$ converges to the truth in both states (G, P) and (G', P), which share the same sampling distribution. So it much that $H_G = H_{G'}$. It follows that, since $\hat{H}$ converges to the truth in state (G', P), it also does in state (G, P), as desired.

To show that property (c) applies to $\hat{H} = F \circ T$, suppose that $\hat{H}$ converges to the truth in a causal state $s = (G, P) \in \mathcal{S}$. So $H_G = F(\mathcal{I}(P))$. Then, by lemma 16, G is minimal to $\mathcal{I}(P)$, so G is minimal to P. By proposition 12, we have:

(i) State s has an open neighborhood $B_\epsilon(s)$ with a sufficiently small radius ϵ such that, for any state $s' = (G', P')$ in that open neighborhood $B_\epsilon(s)$, $G' = G$ and $\mathcal{I}(G) \subseteq \mathcal{I}(P') \subseteq \mathcal{I}(P)$.

Now, recall that, by construction step 1, super test T consists of a test T_σ for each conditional independence statement σ in $\mathcal{I}(\mathcal{V})$ with the convergence properties established in proposition 15. So:

(ii) For each conditional independence statement σ_i in $\mathcal{I}(G)$, which holds everywhere on open ball $B_\epsilon(s)$ by (i), the test T_{σ_i} converges to the correct acceptance of σ_i uniformly on $B_\epsilon(s)$, by clause 1 of proposition 15.

(iii) For each conditional independence statement σ_j in $\mathcal{I}(\mathcal{V}) \smallsetminus \mathcal{I}(P)$, which is violated everywhere on open ball $B_\epsilon(s)$ by (i), there exists a radius $\epsilon_j \leq \epsilon$ such that test T_{σ_j} converges to the correct rejection of σ_j uniformly on $B_{\epsilon_j}(s)$, by clause 2 of proposition 15.

Now, let ϵ' be the minimum of the radius ϵ and the radii ϵ_j constructed in (iii). Then we have:

(iv) Causal state s has an open neighborhood, namely $B_{\epsilon'}(s) \subseteq \mathcal{S}_G$ with $\epsilon' \leq \epsilon$, on which the test T of conditional independence converges uniformly to the correct acceptance of all the conditional independence statements in $\mathcal{I}(G)$ (by (ii)) and the correct rejection of all the conditional independence statements in $\mathcal{I}(\mathcal{V}) \smallsetminus \mathcal{I}(P)$ (by (iii)). That is,

$$\inf_{s' \in B_{\epsilon'}(s)} \mathbb{P}_{s'}\Big(\mathcal{I}(G) \subseteq T(\mathbf{X}_1, \ldots, \mathbf{X}_n) \subseteq \mathcal{I}(P)\Big) \to 1 \quad \text{as} \quad n \to \infty\,.$$

Since $F(\mathcal{I}(P)) = H_G$, we have: $\mathcal{I}(G) \subseteq T(\mathbf{x}_1, \ldots, \mathbf{x}_n) \subseteq \mathcal{I}(P)$ implies $F(T(\mathbf{x}_1, \ldots, \mathbf{x}_n)) = H_G$ (by the second clause of lemma 16). It follows that

$$\inf_{s' \in B_{\epsilon'}(s)} \mathbb{P}_{s'}\Big(F(T(\mathbf{X}_1, \ldots, \mathbf{X}_n)) = H_G\Big) \to 1 \quad \text{as} \quad n \to \infty\,.$$

For every causal state s' in $B_{\epsilon'}(s)$, s' shares the same causal graph G with s, so $H_{s'} = H_G$. Therefore,

$$\inf_{s' \in B_{\epsilon'}(s)} \mathbb{P}_{s'}\Big(F(T(\mathbf{X}_1, \ldots, \mathbf{X}_n)) = H_{s'}\Big) \to 1 \quad \text{as} \quad n \to \infty\,.$$

But $\hat{H} = F \circ T$. So,

$$\inf_{s' \in B_{\epsilon'}(s)} \mathbb{P}_{s'}\Big(\hat{H}(\mathbf{X}_1, \ldots, \mathbf{X}_n) = H_{s'}\Big) \to 1 \quad \text{as} \quad n \to \infty\,,$$

which establishes property (c), as desired. ■

The existence result (clause 2) of theorem 1 follows immediately from the preceding proposition.

Appendix B. Proof of Theorem 2

This appendix is devoted to proving theorem 2. Clause 1.2 follows immediately from lemmas 4 and 5 (in section 6). So it remains to establish clause 1.1 and clause 2.

To establish clause 1.1, let $\hat{H}$ be any learning method for causal learning problem $(\mathcal{V}, \mathcal{S}, \mathcal{H})$ that achieves the joint mode (a)+(b)+(c). Suppose for *reductio* that there is a u-minimal causal state (G, P) in which $\hat{H}$ does not converge to the truth, i.e., does not converge to H_G. Consider the learning method $\hat{H}^*$ that rides on $\hat{H}$ as follows.

DEFINITION OF $\hat{H}^*$: Run a super test T of all the conditional independence statements about $\mathcal{V}$, with the convergence properties established in proposition 15. If T accepts exactly the statements in $\mathcal{I}(P)$ (no more and no less), returns H_G; otherwise, apply $\hat{H}$.

We now show that $\hat{H}^*$ converges to the truth in every causal state in which $\hat{H}$ does. Let (G', P') be any causal state in which $\hat{H}$ converges to the truth. By clause 1.2 (which has been established), (G', P') is a minimal causal state. To show that $\hat{H}^*$ converges to the truth in state (G', P'), discuss two exhaustive cases: either $\mathcal{I}(P') = \mathcal{I}(P)$, or not. Case 1: suppose that $\mathcal{I}(P') = \mathcal{I}(P)$. Then $\mathcal{I}(G') = \mathcal{I}(G)$, for three reasons: first, (G, P) is u-minimal; second, (G', P') is minimal; and third, $\mathcal{I}(P') = \mathcal{I}(P)$. Since $\mathcal{I}(G') = \mathcal{I}(G)$, we have that $H_{G'} = H_G$. Note that T has the convergence properties established in proposition 15; so, in state (G', P'), T converges to the acceptance of all and only the statements in $\mathcal{I}(P')$, which is identical to $\mathcal{I}(P)$. So $\hat{H}^*$ converges to hypothesis H_G in state (G', P'). But $H_G = H_{G'}$. So $\hat{H}^*$ converges to the truth $H_{G'}$ in state (G', P'). Now turn to case 2: suppose that $\mathcal{I}(P') \neq \mathcal{I}(P)$. So, in state (G', P'), T converges to exactly the statements in $\mathcal{I}(P')$, and hence it is not the case that T converges to exactly the statements in $\mathcal{I}(P)$. So, in state (G', P'), $\hat{H}^*$ converges to whatever $\hat{H}$ converges to. With the above discussion of the two exhaustive cases, it follows that $\hat{H}^*$ converges to the truth in every causal state in which $\hat{H}$ does. Moreover, thanks to T, $\hat{H}^*$ converges to the truth in state (G, P), in which $\hat{H}$ does not by hypothesis. Therefore, $\hat{H}$ does not achieve convergence to truth on a maximal domain—contradiction. This establishes clause 1.1.

To establish clause 2, consider any causal state (G_1, P) that is minimal but not u-minimal. Since it is not u-minimal, there exists G_2 such that $H_{G_1} \neq H_{G_2}$ and (G_2, P) is also a minimal causal state. By lemma 18 and proposition 19, there is a learning method $\hat{H}_1$ that achieves the joint mode (a)+(b)+(c) and converges to the truth in (G_1, P), and a learning method $\hat{H}_2$ that achieves the joint mode (a)+(b)+(c) and converges to the truth in (G_2, P). Since $H_{G_1} \neq H_{G_2}$, $\hat{H}_2$ does not converge to the truth in (G_1, P). Clause 2 follows.

Appendix C. An Illustrated Explanation of Why Clause 3 of Theorem 1 Holds

Recall the example illustrated in figure 2; for ease of reference, it is illustrated below in the simplified figure 6. Note that the left, cubic state space embeds a trapezoidal plane, which is an identical

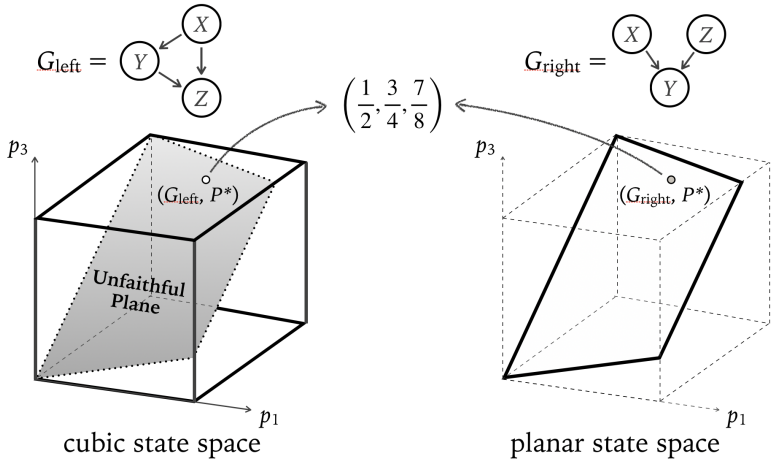


Figure 6: violation of adherently locally uniform convergence

copy of the planar state space on the right. The left trapezoid contains all and only the unfaithful causal states in the left cubic state space; so call it the **unfaithful plane**, as indicated in figure 6. Every causal state on the left, embedded trapezoid shares an identical joint distribution with a corresponding causal state on the right, planar state space. For any such pair of causal states, the convergence property has to be sacrificed in at least of the two. The standard design practice would sacrifice the convergence property on the left trapezoid. But consider the alternative proposal that makes sacrifices in accordance with the standard practice except that, for the distribution P^* parametrized by $(p_1, p_2, p_3) = \left(\frac{1}{2}, \frac{3}{4}, \frac{7}{8}\right)$, sacrifices are made in the right, faithful causal state (G_{right}, P^*) instead of the left, unfaithful causal state (G_{left}, P^*). So, on this alternative proposal, the shaded areas in figure 6 are the places where sacrifices are made: a shaded point on the right, together with a shaded, punched plane on the left. On this alternative proposal, the convergence property is secured in the left causal state (G_{left}, P^*) but sacrificed in some causal states that are *arbitrarily close* to that causal state, which leads to a violation of adherently locally uniform convergence.

To avoid such a violation, one might try to secure the convergence property not just in the left causal state (G_{left}, P^*) but in all of its nearby states, as depicted by the open disc on the left side of figure 7. (The shaded areas are still understood as the places where sacrifices are made.) But doing

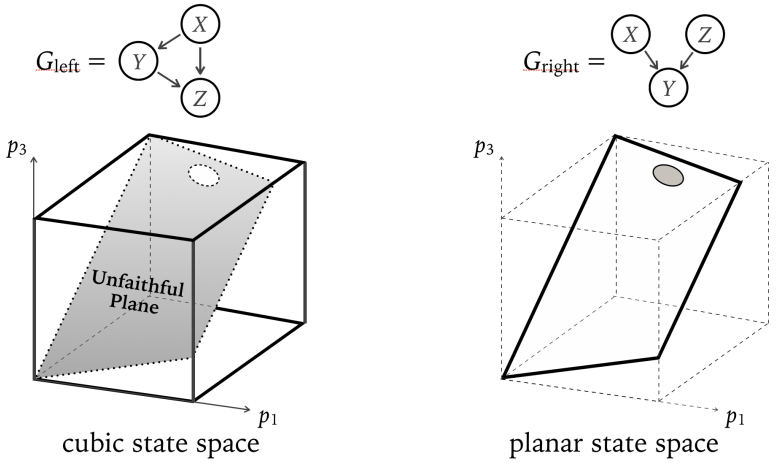


Figure 7: violation of almost everywhere convergence

so would force the convergence property to be sacrificed on the corresponding disc on the right, planar state space, which leads to a violation of almost everywhere convergence.

So, there is only one way to avoid both the two kinds of violations depicted in figures 6 and 7: given any causal state (G_{left}, P) on the left, unfaithful plane and the corresponding causal state (G_{right}, P) in the right, planar state space, the convergence property has to be sacrificed in the left, unfaithful one.

This allows for the possibility of converging to the truth in any faithful causal state. And this possibility can be forced into a reality by requiring a maximal domain of convergence.

Appendix D. Some More Details of the Example in Section 2

Given the background assumption of the example in section 2, there are two possible causal structures on the table with three parameters p_1, p_2, and p_3 whose values are unknown. For each of those two causal structures, the conditional probability of every effect given its immediate causes can be expressed by the three parameters, as indicated in figure 8.

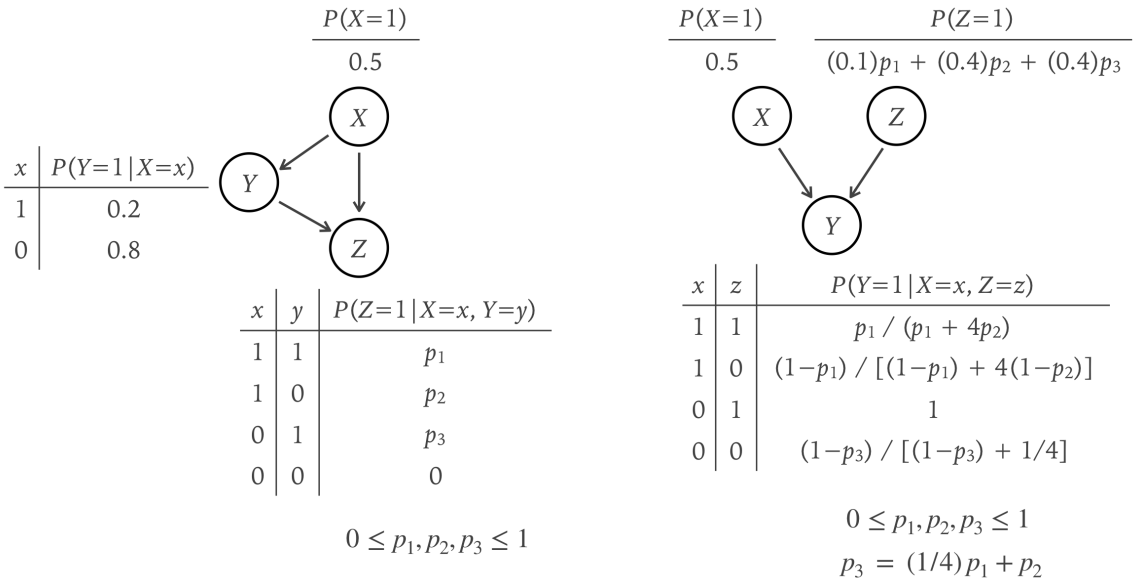


Figure 8: Conditional probability tables

When it is said that the same joint distribution is shared, what is actually meant is only that the same joint distribution is shared *in the absence of manipulation*. To illustrate, consider the two CBNs $\left(G_{\text{left}}, P^*\right)$ and $\left(G_{\text{right}}, P^*\right)$ that share the same joint distribution P^* parametrized by $(p_1, p_2, p_3) = (\frac{1}{2}, \frac{3}{4}, \frac{7}{8})$. Also consider the manipulation that forces $Y = 0$. If the true CBN is the right one, the manipulation $Y = 0$ is only a manipulation of an effect rather than a cause (see the right causal graph G_{right}); so the distribution of Z would remain the same were this manipulation applied—in particular, the probability of $Z = 0$ would remain at 30%. But the same manipulation would raise the probability of $Z = 0$ from 30% to 62.5% if instead the true CBN is the one on the left $\left(G_{\text{left}}, (\frac{1}{2}, \frac{3}{4}, \frac{7}{8})\right)$. Indeed, in this case, the manipulation $Y = 0$ is a manipulation of a cause of Z. Therefore, whether the true CBN is the one on the left or on the right makes an important difference—at least for those who are thinking about manipulating Y in order to change Z. So it would be great if there exists a learning method that can distinguish between those two CBNs. Unfortunately, there exists no such learning method if the available data are non-experimental (i.e., collected without any manipulation of the true, unknown CBN), as explained in section 3.