The Nonidentifiability Problem in Causal Discovery

Hanti Lin
Department of Philosophy
University of California, Davis
Davis, CA 95616
ika@ucdavis.edu

Abstract

Lin and Zhang's (2020) learning-theoretic approach to the problem of nonidentifiability proceeds without making the causal faithfulness assumption (or any of its weaker variants) but still manages to explain why we should follow this standard design principle: when there is no pre-existing reason against faithfulness, develop and adopt causal learning algorithms that converge to the truth for at least all faithful causal Bayes nets. But their result has three limitations: First, it concerns only categorical variables. Second, it does not take care of other cases of nonidentifiability such as those that have motivated the work on LiNGAM (Shimizu et al. 2006). Third, they adopt a conception of "almost all" or "almost everywhere" that is topological, and it is not (yet) clear why it has to be topological rather than measure-theoretic. This paper aims to remove those three limitations.

1 Introduction

Suppose that we are considering two directed acyclic graphs as possible causal structures, such as

$$\begin{aligned} G_1 &= \{X \to Y \leftarrow Z\}\,, \\ G_2 &= \{X \to Y \to Z,\ X \to Z\}\,. \end{aligned}$$

We are wondering which one is true, and can only infer an answer from purely non-experimental, non-temporal data. We assume that the true causal model is a causal Bayes net on one of those two causal structures, and that data are generated by IID observations. And we assume no more than that. Then it is well-known that every learning algorithm for this learning task is doomed to perform badly in some cases, no matter how large the sample size may be. Here is why. Suppose that God has told us that the true probability distribution is P, which does for us all the work about statistical inference. Suppose further that P turns out to make X and Z independent. Then P can be paired either with G_1 or with G_2 to form a causal Bayes net: (G_1, P) and (G_2, P). But then which of the two causal Bayes net is the true one? Which causal structure is true? There is no way to tell (given the restriction to purely non-experimental, non-temporal data): at any sample size, if an algorithm has a high chance $1 - \epsilon$ of getting it right with respect to one of the two Bayes nets, then that algorithm must have a low chance no more than ϵ of getting it right with respect to the other Bayes net. This is because those two Bayes nets share the same joint distribution and, hence, the same sampling distribution under the IID assumption. So, there is no way to identify the true one between those two, and thus there arises the problem of *nonidentifiability*.

The standard approach to this problem relies on making an assumption to rule out all but one of the causal Bayes nets that share the same joint distribution—such as ruling out (G_1, P) or (G_2, P). But which to rule out? The standard answer is to rule out (G_2, P) and keep (G_1, P)—to keep any Bayes net that is so-called faithful (Spirtes et al. 2000). But why make this assumption? That is, why should we rule out possibilities in that way? Why not make another assumption? Why not, for example, swap the roles of those two Bayes nets and keep everything else the same: rule out the

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

faithful one (G_1, P) and keep instead its unfaithful counterpart (G_2, P), and then keep all the other faithful Bayes nets? Furthermore, why make an assumption to rule out any Bayes net at all, if there is no pre-existing reason to make such an assumption?

In reply to the above questions Lin (2019) and Lin and Zhang (2020) develop an approach to the problem of nonidentifiability in causal discovery that does not rely on ruling out unfaithful causal Bayes nets, but still explain why a good learning algorithm should follow this standard design principle: converge to the truth for at least every faithful Bayes net, as long as we have no pre-existing reason to rule out any faithful one. (To be sure, when we do have pre-existing reason to rule out the faithful ones, it is pointless to design a learning algorithm that converges to the truth for the faithful ones.) Their idea is that, although nonidentifiability makes it impossible for a learning algorithm to achieve convergence to the truth for every possible Bayes net under consideration, we can still look for the best achievable mode of convergence and show that, to achieve the best achievable, a learning algorithm has no alternative but to follow the standard design principle mentioned above: securing convergence to the truth for at least every faithful Bayes net.

To clarify, the above result is *not* meant to justify acceptance of the faithfulness assumption; it is *not* meant to justify ruling out any unfaithful Bayes net. The above result is meant to address a problem that arises in the case in which we make very weak assumptions, so much so that we drop the faithfulness assumption and any of its variants, and hence we have to design and adopt a learning algorithm that sacrifices the property of convergence for *some* Bayes nets on the table. Some; but *which*? That's the question.

While I believe that Lin (2019) and Lin and Zhang (2020) are on the right track, their results are limited in at least three aspects: First, it concerns only categorical variables and misses real-valued variables. Second, although it takes care of the kind of nonidentifiability that arises in the work on faithfulness, it has not (yet) said anything about another important kind of nonidentifiability that motivated the work on LiNGAM (Shimizu et al. 2006). Third, they point out that, when convergence for all Bayes nets is unachievable, we should see whether it is possible to achieve convergence for almost all. While this idea seems to be on the right track, they adopt a conception of "almost all" or "almost everywhere" that is topological, in stark contrast with the more familiar, measure-theoretic approach often adopted in causal discovery (Spirtes et al. 2000, Meek 1995). It is not (yet) clear why the topological approach is preferred.

This paper aims to remove those three limitations and extend Lin's (2019) and Lin and Zhang's (2020) main results. In fact, this paper aims to do more. An important guiding principle from them is that, when tackling any learning problem, we should design learning algorithms that achieve the best achievable mode of convergence—best achievable with respect to the learning problem in question. If this is right, then there is a need to explore various new modes of convergence. While the previous works take an initial step, this paper aims to start a more systematic exploration of much more modes of convergence, as depicted in figure 1: All the modes of convergence (1)-(7) in this figure will be defined in due course before I state any result about this figure. Each line in the diagram (solid or dotted) is understood to assert an implication: the higher implies the lower. While the previous works only study modes (1) and (3), I set out to systematically study all modes in this figure.

2 Preliminaries

Consider a finite set of variables, $\mathcal{V} = \{X_1, X_2, \dots, X_n\}$. A possible **causal structure** over those variables is represented by a directed acyclic graph on $\mathcal{V}$, written $G = (\mathcal{V}, \to)$, where the binary relation $X_i \to X_j$ 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 connected in such a way 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) Bayes net**. The Markov condition, as the defining condition of causal Bayes nets, is often taken for granted as a necessary connection between causal graphs and joint distributions—between causation and probability.

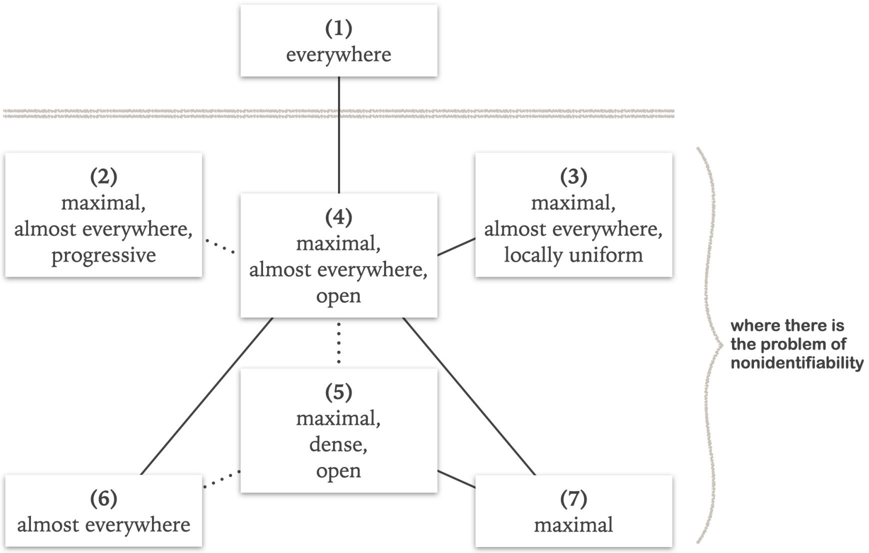

Figure 1: Modes of convergence, ordered by implication

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

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

Now, consider the following stronger condition:

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

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 causal Bayes net (G, P) satisfies the **faithfulness condition**. With respect to an **unfaithful** causal Bayes net (G, P), at least one conditional independence statement σ happens to hold (i.e., $\sigma \in \mathcal{I}(P)$) even though it is not required to hold by the Markov condition (i.e., $\sigma \notin \mathcal{I}(G)$).

A **causal learning problem** is represented by a triple $(\mathcal{O}, \mathcal{S}, \mathcal{H})$ whose components meet the following conditions:

1. (Observable Variable) $\mathcal{O}$ is a set of variables; it is understood to contains all and only the variables that are observable in our inquiry.
2. (State Space) $\mathcal{S}$ is a set of causal Bayes nets over the variables in $\mathcal{O}$ and possibly some more variables. It is interpreted as the state space that contains all and only the possible states of the world that are compatible with whatever taken to be the background assumption, i.e., whatever we are happy to take for granted in our inquiry. When this interpretation needs to be emphasized, the Bayes nets in $\mathcal{S}$ will be called (possible) **causal states** (of the world).
3. (Hypothesis) $\mathcal{H}$ is a set of competing hypotheses about causal structures, assumed to be mutually exclusive and jointly exhaustive given $\mathcal{S}$. So, in every causal Bayes net in $\mathcal{S}$, exactly one hypothesis in $\mathcal{H}$ is true; two causal Bayes nets with the same causal graph shares the same true hypothesis in $\mathcal{H}$.

Let $\mathrm{Val}(\mathcal{O})$ be the set of the assignments of values to all the variables in $\mathcal{O}$. A **data sequence** (of sample size n) is a sequence of such assignments of values (of length n). A **learning method** M is a function from any data sequence (of a finite length) to one of the hypotheses in $\mathcal{H}$.

Learning methods are to be evaluated on the basis of their performances in each of the causal states in $\mathcal{S}$, and performances are identified with the probabilities of outputting the truth. This idea can be made more precise as follows. In each causal state $s = (G, P) \in \mathcal{S}$, there exists a uniquely true hypothesis, denoted by $H_s \in \mathcal{H}$, and the true probability distribution P induces a sampling distribution Pr_s^n over the data sequences of sample size n under the IID assumption. Let's think about the probability that learning method M would output the true hypothesis H_s if it were given a sample of size n in causal state s; denote this probability by

$$\mathsf{Pr}_s^n(M \text{ outputs the truth}) \quad =_{\text{def}} \quad \mathsf{Pr}_s^n\Big(\big\{\vec{v} \in (\text{Val}(\mathcal{O}))^n : M(\vec{v}) = H_s\big\}\Big)\,.$$

With respect to a causal learning problem $(\mathcal{O}, \mathcal{S}, \mathcal{H})$, a learning method M is said to be **pointwise consistent** if

$$\forall s \in \mathcal{S},\ \lim_{n\to\infty} \mathsf{Pr}_s^n(M \text{ outputs the truth}) \ = \ 1\,.$$

The problem of nonidentifiability arises in non-experimental causal discovery when the state space $\mathcal{S}$ contains two causal Bayes nets $s = (G, P)$ and $s' = (G', P)$ such that they share the same joint distribution P over the observable variables (so $\mathsf{Pr}_s^n = \mathsf{Pr}_{s'}^n$ for each n) but they make distinct hypotheses true (so $H_s \neq H_{s'}$). In this case, pointwise consistency is too demanding to be achievable. Following Lin and Zhang (2020), let's explore new modes of convergence that are not as demanding but still desirable.

3 Modes of Convergence

With respect to a causal learning problem $(\mathcal{O}, \mathcal{S}, \mathcal{H})$, every learning method M has its own domain of convergence (to the truth), written $\mathsf{DC}(M)$ and defined by

$$\mathsf{DC}(M) \quad = \quad \Big\{s \in \mathcal{S} : \lim_{n\to\infty} \mathsf{Pr}_s^n(M \text{ outputs the truth}) = 1\Big\}\ .$$

Pointwise consistency means convergence to the truth **everywhere** in the given state space $\mathcal{S}$:

$$(\text{EVERYWHERE}) \qquad \mathsf{DC}(M) \quad = \quad \mathcal{S}\,. \tag{1}$$

When everywhere convergence is too high a standard to be achievable, what would still be great to have? Here are some informal ideas, to be rigorously defined soon. It would still be great to have a learning method M that has a "maximal" domain of convergence—a domain of convergence that cannot be extended further. It would also be great if the domain of convergence $\mathsf{DC}(M)$ extends so pervasively that it is at least "dense" in the state space $\mathcal{S}$. Moreover, it would be great if M provides a kind of stable learning: whenever M converges to the truth in a possible state in $\mathcal{S}$, M does, too, in nearby states in $\mathcal{S}$. This means that the domain of convergence $\mathsf{DC}(M)$ is "open" in $\mathcal{S}$. The openness condition seems to be a desideratum: If the learning method in use converges to the truth for the true causal Bayes net, then this method would continue to converge to the truth even if the parameters of the true causal Bayes net were perturbed by some causal factors that have not been explicitly considered—as long as the perturbations were sufficiently small.

So, pending a definition of the topology in use, we have the following modes of convergence. With respect to a causal learning problem $(\mathcal{O}, \mathcal{S}, \mathcal{H})$, a learning method M is said to have a domain of convergence that is maximal/dense/open if $\mathsf{DC}(M)$ satisfies condition (2)/(3)/(4), respectively:

$$(\text{MAXIMAL}) \qquad \mathsf{DC}(M) \quad \subsetneq \quad \mathsf{DC}(M') \text{ for no } M'\,. \tag{2}$$

$$(\text{DENSE}) \qquad \mathsf{DC}(M) \quad = \quad \text{a dense subset of } \mathcal{S}\,. \tag{3}$$

$$(\text{OPEN}) \qquad \mathsf{DC}(M) \quad = \quad \text{an open subset of } \mathcal{S}\,. \tag{4}$$

It remains to define and adopt a reasonable topology on the state space $\mathcal{S}$. Consider a quite standard metric on probability measures, the **total variation metric**, 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 that occur in $\mathcal{S}$ (any metric would do). 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:

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

where the radius ϵ is required to be positive. This turns the state space $\mathcal{S}$ into a topological space: A subset of $\mathcal{S}$ is **open** if it is a union of open balls; it is **dense** if it has a nonempty overlap with every open ball/ball. Note that the specific distance functions used to define the open sets are inessential to this paper; it is the open sets that are essential.

Openness captures a kind of stable learning, as mentioned above. In a subsequent section, we will explore stronger and more desirable kinds of stable learning. Similarly, denseness is often taken to be a necessary condition for a set to cover "almost everywhere". A variety of interpretations of "almost everywhere", either topological or measure-theoretic, will be compared in a subsequent section. For now, it suffices to keep in mind that a maximal, dense, open domain of convergence is one of the minimal qualifications for making a good learning method. Here is the point: when we require at least maximality, denseness, and openness, we are already in a position to solve some important instances of the nonidentifiability problem, as we will see in the next two sections.

4 Application I: Doing without Assuming Faithfulness

While Lin and Zhang's (2020) main theorem applies to categorical variables, the following result extends it to cover the case of real-valued variables.

Some more definitions are needed. Two causal graphs are **Markov equivalent** if they entail exactly the same conditional independence statements. Let G be a causal graph on $\mathcal{O}$; H_G be the hypothesis saying that the true causal graph is Markov equivalent to G. Hypotheses of that form are called **Markov equivalence hypothesis** about $\mathcal{O}$.

Hold a causal graph G fixed, and let $\mathsf{PA}(X_i)$ denote the set of the parents of variable X_i. Consider two arbitrary Bayes nets (G, P) and (G, Q) that share G as their common graph. Then we have the (familiar) factorization formulas:

$$\begin{aligned} P(X_1, \ldots, X_n) &= \prod_{i=1}^{n} P\big(X_i \mid \mathsf{PA}(X_i)\big)\,, \\ Q(X_1, \ldots, X_n) &= \prod_{i=1}^{n} Q\big(X_i \mid \mathsf{PA}(X_i)\big)\,. \end{aligned}$$

The following constructs a kind of mixture of those two distributions. First, choose a real number $\alpha \in [0, 1]$ as a mixing coefficient. For each variable X_i, grab the two corresponding multiplicative terms $P\big(X_i \mid \mathsf{PA}(X_i)$ and $Q\big(X_i \mid \mathsf{PA}(X_i)$ to form a mixture $\alpha\, P\big(X_i \mid \mathsf{PA}(X_i)\big) \,+\, (1{-}\alpha)\, Q\big(X_i \mid \mathsf{PA}(X_i)\big)$. Then multiply all such mixtures to obtain a new distribution D:

$$D(X_1, \ldots, X_n) \;=\; \prod_{i=1}^{n} \Big[\alpha\, P\big(X_i \mid \mathsf{PA}(X_i)\big) \,+\, (1{-}\alpha)\, Q\big(X_i \mid \mathsf{PA}(X_i)\big)\Big]\,.$$

This is called a **Markov mixture** of distributions P and Q with respect to the common causal graph G, denoted by $\alpha P {+_{G}} (1{-}\alpha) Q$. It is routine to show that is also Markov to G. So $(G, \alpha P {+_{G}} (1{-}\alpha) Q)$ is a Bayes net, called a **Markov mixture** of Bayes nets (G, P) and (G, Q). Note that a Markov mixture is defined only when there is a common graph.

A state space $\mathcal{S}$ of Bayes nets is said to be **closed under Markov mixtures** if, whenever it contains two Bayes nets that share a common graph, it also contains all of their Markov mixtures. State space $\mathcal{S}$ is said to be **closed under graph replacements** if, whenever it contains a Bayes net (G, P), then it also contains (G', P) for any G' that is Markov to P. This last closure property suggests that the background assumption turns out to be too weak to rule out unfaithful Bayes nets. When $\mathcal{S}$ contains enough Bayes nets to have the two closure properties just defined, say that it is **inclusive**.

Sate space $\mathcal{S}$ is said to be **regular** if every conditional independence hypothesis about $\{X_1, \ldots, X_n\}$ has a test that is consistent with respect to all joint distributions that appear in $\mathcal{S}$ (that is, for all P such that $(G, P) \in \mathcal{S}$ for some G).

Then we have the first main result:

Theorem 1. *Suppose that $(\mathcal{O}, \mathcal{S}, \mathcal{H})$ is a causal learning problem that meets the following conditions:*

- *$\mathcal{O}$ is a finite set of real-valued variables (taken to be the observable variables).*

- *$\mathcal{S}$ is an inclusive and regular space of Bayes nets over $\mathcal{O}$ (namely, over just the observable variables).*

- *$\mathcal{H}$ contains exactly the Markov equivalence hypotheses about $\mathcal{O}$.*

Then, with respect to this learning problem, we have:

1. (POSSIBILITY RESULT) *There exists a learning method whose domain of convergence is maximal, dense, and open.*

2. (NECESSITY RESULT) *For any learning method M whose domain of convergence is maximal, dense, and open, M converges to the truth in every faithful causal state in $\mathcal{S}$.*

5 Application II: Doing without Assuming Faithfulness + LiNGAM

What if we want to learn a more fine-grained truth—more fine-grained than the Markov equivalence hypotheses—without making any additional assumption? Unfortunately the result is negative. For simplicity, from now on we consider only the bivariate case in which only two variables X and Y are of interest. So let $\mathcal{O} = \{X, Y\}$.

Theorem 2 (Impossibility Result)**.** *Suppose that $(\mathcal{O}, \mathcal{S}, \mathcal{H})$ is a causal learning problem that meets the following conditions:*

- *$\mathcal{O}$ contains exactly two real-valued variables X and Y.*

- *$\mathcal{S}$ contains only Bayes nets on $\mathcal{O} = \{X, Y\}$ and is closed under graph replacements.*

- *$\mathcal{H}$ contains exactly these three hypotheses: (i) "X causes Y", (ii) "Y causes X", and (iii) "neither causes the other" (so it is more fine-grained than the set of the Markov equivalence hypotheses).*

Then, with respect to this learning problem, there exists no learning method whose domain of convergence is both dense and open.

There is still some hope for learning about the causal connection between X and Y if the background assumption (as represented by $\mathcal{S}$) is strong enough to imply that the true Bayes net is a so-called linear acyclic model. In general, a Bayes net (G, P) is said to be a **linear acyclic model** if, for any variable V_i therein, the conditional distribution of V_i given its parents takes the form

$$V_i \;=\; \sum_{j \in J} c_{ij} U_{ij} + \varepsilon_i \, ,$$

where the variables in $\{U_{ij}\}_{j \in J}$ are exactly the parents of V_i, the coefficients c_{ij} are constants, and ε_i is a random error term, with the property that the random variables on the right hand side (ε_i and U_{ij} with $j \in J$) are mutually independent. As a convention, in the case where V_i has no parent, J is empty, U_{ij} does not exist, and the above equation simplifies to

$$V_i \;=\; 0 + \varepsilon_i \, .$$

A linear acyclic model is said to be **Gaussian** if, for each variable V_i therein, its error term ε_i is Gaussian; otherwise, it is said to be **non-Gaussian**. A linear acyclic model is said to be **degenerate** if at least one variable V_i therein has a parent U_{ij} with coefficient $c_{ij} = 0$; otherwise, it is said to be **non-degenerate**.

Now let's return to the case of bivariate Bayes nets for simplicity. Suppose that somehow we are happy to take for granted (i.e., make the background assumption) that the true Bayes net on $\{X, Y\}$ is a linear acyclic model. And we are wondering whether any of the two variables causes the other and, if so, which does. To see how nonidentifiability arises in this setting, note that the bivariate linear acyclic models (G, P) can be categorized as follows:

Case A Suppose that linear acyclic model (G, P) is Gaussian, non-degenerate, and with a causal edge $X \to Y$.

In this case, the conditional distribution of Y given X takes the form $Y = cX + \varepsilon'$ with $X \perp\!\!\!\perp \varepsilon'$ and $c \neq 0$, and that either X or ε' is Gaussian. Then (G, P) has a counterpart

(G', P) that shares the same joint distribution P, has a reversed graph $G' = \{Y \to X\}$, and is in itself a Gaussian linear acyclic model (Shimizu et al. 2006). So both (G, P) and (G', P) are on the table. So here arises nonidentifiability, which leads to the next case.

Case A' Suppose that linear acyclic model (G, P) is Gaussian, non-degenerate, and with a causal edge $Y \to X$.

So this is same as case A except that the roles of X and Y are exchanged. This case is the nonidentifiable counterpart of case A.

Case B Suppose that linear acyclic model (G_{empty}, P) is degenerate without a causal edge, i.e, $X \not\to Y$ and $Y \not\to X$ (so G_{empty} is the empty graph).

In this case, X and Y are independent (thanks to the Markov condition) and the model is faithful.

Case B' Suppose that linear acyclic model (G, P) is degenerate with a causal edge $X \to Y$ or $Y \to X$.

In this degenerate case, if $G = \{X \to Y\}$, then $X = \varepsilon$, $Y = cX + \varepsilon'$ with $c = 0$, and X are ε' are independent. This implies the independence between X and Y, despite the fact that it is not entailed by the graph $G = \{X \to Y\}$. So (G, P) is *unfaithful*. It has a nonidentifiable counterpart: the faithful linear acyclic model (G_{empty}, P), which is the previous case. So here arises nonidentifiability.

Case C Suppose that linear acyclic model (G, P) is in none of the above cases, i.e., it is non-Gaussian, non-degenerate, and has a causal edge $X \to Y$ or $Y \to X$.

In this case, (G, P) as a Bayes net does have a counterpart (G_{reversed}, P) that shares the same joint distribution and reverses the causal edge. But it can be shown that, although this counterpart (G_{reversed}, P) is a Bayes net nonetheless, it fails to be a linear acyclic model and, hence, is ruled out by what we have taken for granted (Shimizu et al. 2006). So this case does not have a nonidentifiable counterpart on the table.

So, the instances of case A are nonidentifiable counterparts of the instances of case A'; and similarly for cases B and B'. In this situation, there is an often-used assumption that rules out non-identifiability by leaving open just the instances in case C: it is the LiNGAM assumption, stating that the true Bayes net is a non-degenerate *li*near *n*on-*g*aussian *a*cyclic *m*odel. But there are infinitely many other possible assumptions to make in order to rule out nonidentifiability. Which of such assumptions to make?

And what if we are to design learning algorithm *without* making any such assumption? Then there is actually a reason to secure the convergence property at least for the faithful case B and the LiNGAM case C and sacrifice it for the unfaithful case B' Again, the idea of a maximal, dense, and open domain of convergence will play an important role:

Theorem 3. *Suppose that $(\mathcal{O}, \mathcal{S}, \mathcal{H})$ is a causal learning problem that meets the following conditions:*

- *$\mathcal{O}$ contains exactly two real-valued variables X and Y.*
- *$\mathcal{S}$ contains only linear acyclic models on $\mathcal{O} = \{X, Y\}$, is closed under graph replacements; there is a consistent test of the independence between X and Y and a consistent test of Gaussianity for all joint distributions that appear in $\mathcal{S}$.*
- *$\mathcal{H}$ contains exactly these three hypotheses: (i) "X causes Y", (ii) "Y causes X", and (iii) "neither causes the other".*

Then, with respect to the above learning problem, we have:

1. (POSSIBILITY RESULT) *There exists a learning method whose domain of convergence is maximal, dense, and open.*
2. (NECESSITY RESULT) *For any learning method M whose domain of convergence is maximal, M converges to the truth in at least all instances of case C, i.e., all non-degenerate non-Gaussian models in $\mathcal{S}$.*
3. (NECESSITY RESULT) *For any learning method M whose domain of convergence is maximal, dense, and open, then M converges to the truth in all instances of case C, in all (faithful) instances of case B, and in no (unfaithful) instances of case B'.*

6 More Modes of Convergence

If the goal is to design a learning algorithm that achieves the best achievable mode of convergence, we should try to explore more modes of convergence and do it systematically.

6.1 Be Stable: Openness and Beyond

An open domain of convergence represents a mode of stable learning, as explained above. But we should try to achieve something better if possible. The following presents two stronger, more desirable modes of stable learning.

Lin and Zhang (2020) introduce a kind of locally uniform convergence: A learning method M is said to converge to the truth with **local uniformity** if, for any causal state $s \in \mathcal{S}$, if M converges to the truth in s, then M converges to the truth uniformly on some open neighborhood $B_\epsilon(s)$ of s in the state space $\mathcal{S}$. This means that, in every causal state in which the learning method converges to the truth, the probability of error can be made not just low but *stably* low. Then we have:

Lemma 1 (Lin and Zhang 2020)**.** *With respect to any causal learning problem, local uniformity implies an open domain of convergence.*

Genin (2018) introduces a kind of near-monotonic convergence that captures the idea of making steady progress. Let $\alpha > 0$. A learning method M is said to converge to the truth α-**progressively** if, for any causal state $s \in \mathcal{S}$, if M converges to the truth in s, the probability for M to output the truth never drops by more than α as the sample size increases; that is, for any $s \in \mathsf{DC}(M)$, any positive integers n and k,

$$\mathsf{Pr}_s^{n+k}(M \text{ outputs the truth}) \quad \geq \quad \mathsf{Pr}_s^{n}(M \text{ outputs the truth}) - \alpha$$

Lemma 2. *With respect to any causal learning problem, for any learning method that converges to the truth α-progressively, its domain of convergence, if maximal and dense, must be open.*

6.2 Be Almost Everywhere: Denseness and Beyond

It would be great to have convergence to the truth (even though that may not be the only good thing to have). But we are confronted with severe uncertainty, with a wide range of possible cases—so wide that it is impossible to extend the domain of convergence to cover the entire space $\mathcal{S}$ of the possible states under consideration. But it would still be great if the domain of convergence could be extended to cover "almost everywhere" in the state space $\mathcal{S}$. Mathematicians have developed many conceptions of "almost everywhere". So which one should we use?

There are two broad approaches: topological and measure-theoretic. The measure-theoretic one is familiar: if the state space is finite dimensional, when it is convenient to interpret "almost everywhere" as "everywhere but a set of Lebesgue measure zero". This conception of "almost everywhere" has been most popular in the causal discovery literature, thanks to Spirtes et al. (2000). But it has a limitation: it applies only when the state space is finite dimensional, such as when the variables are all categorical (Meek 1995), or when the variables are real-valued but assumed to take a strong parametric form (Spirtes et al. 2000). When we consider a nonparametric setting (as we have done for theorem 1) or a semiparametric setting (as we have done for theorem 3), the state space is too big to be finite dimensional, and thus too big to allow for a straightforward analogue of the Lebesgue measure. There is a not-so-straightforward analogue: mathematicians have developed the concept of so-called *prevalent* subsets as a measure-theoretic concept of "almost everywhere" in the infinite dimensional setting (Hunt 1994), and it might one day be applied to causal discovery. That said, it seems much easier to employ a topological approach, to which I turn.

The topological approach has a long history in mathematics but was adopted in causal discovery only recently (Lin and Zhang 2020): taking "almost everywhere" as "everywhere but a nowhere dense set". A nowhere dense set is a topologically small set, incredibly full of holes: to construct a **nowhere dense** set is, by definition, to follow these two steps: (i) given a state space $\mathcal{S}$ equipped with a topology, remove an open set within *every* open neighborhood of *every* point in $\mathcal{S}$; (ii) stop there or even remove some more points. This approach to the concept of "almost everywhere" has an advantage: it gives a uniform treatment for the finite dimensional setting as well as the more general settings. As long as the state space $\mathcal{S}$ in question comes with a natural topology, we are good to go.

The above discussion seems to suggest that the topological approach has the advantage of providing a more uniform treatment while staying with relatively simple mathematics—at least for now.

But I would also like to argue that there is no urgent need to choose between those approaches to "almost everywhere". Here is the idea. We have seen that it is (i) maximality, (ii) openness, and (iii) denseness that work together as an essential package to deal with the problem of nonidentifiability. The first of the three elements, maximality, is obviously desirable and can be motivated independently without considering almost everywhere convergence. The second, openness, is desirable insofar as stable learning is desirable, so it can also be motivated independently without considering almost everywhere convergence (as discussed in the previous subsection). So, the work that really needs to be done by the concept of "almost everywhere" is not much: we only need it to imply denseness to complete the tripartite package of maximality, openness, and denseness. Fortunately, it is well known that denseness is indeed entailed by *each* of the three interpretations of "almost everywhere" mentioned above (i.e., the Lebesgue interpretation, the prevalence interpretation, the nowhere denseness interpretation). So there is no urgent need to choose among the three. I thereby propose the following axiom as a condition of adequacy for good interpretations of "almost everywhere":

Axiom. *A domain of convergence covers almost everywhere only if it is dense.*

Let me defend this axiom against a potential worry. It might be worried that there is a topological interpretation of "almost everywhere" that fails to imply denseness. Instead of taking a set that covers "almost everywhere" as a set that covers everywhere except a nowhere dense set, we can *lower* the standard by taking it as a **comeager** set, i.e. a set that covers everywhere except a countable union of nowhere dense sets. In reply, I want to ask: Why lower the standard? We should achieve more if we can. So, don't sacrifice the convergence property on a countable union of nowhere sense sets, if you only need to sacrifice it on just one nowhere sense set, which implies a dense domain of convergence. So we should worry about the above axiom only when there is an interesting learning problem that requires us to lower the standard for "almost everywhere" because the higher standard is unachievable. Moreover, even when we have to settle with the lower standard in terms of comeager sets, we need to worry about the above axiom only when there really is an interesting learning problem for which the state space has a comeager set that fails to be dense. For many spaces, comeager sets must be dense sets as well. Examples abound, including every complete metric/metrizable space (thanks to the Baire category theorem). Finally, even when there really is an interesting learning problem for which the state space has a comeager set that fails to be dense, it is not clear that "comeager" really serves as a good interpretation of "almost everywhere" for that state space—perhaps some other concept works better. I submit that only a concept that implies denseness works well for interpreting "almost everywhere", following the above axiom.

6.3 Big Picture

We have discussed many modes of convergence. Some interesting combinations of those modes are presented in figure 1 (which has appeared in the introductory section). Thanks to the previous two lemmas 1 and 2 and the axiom just defended, we have an immediate result for the modes of convergence in figure 1:

Theorem 4. *In figure 1, understand each line as an implication pointing downward. Every solid-line implication holds. Dotted-line implications* $(2) \Rightarrow (4)$ *and* $(4) \Rightarrow (5)$ *hold under the axiom that "almost everywhere" implies "dense". Dotted-line implication* $(5) \Rightarrow (6)$ *holds if "almost everywhere" is identified with "everywhere except a nowhere dense set" (which implies the above axiom).*

7 Conclusion

A core idea of learning theory is that, when tackling a learning problem, a good learning algorithm should achieve a highest achievable mode convergence (to the desired learning target). I suggest that the essential mode of convergence for solving the nonidentifiability problem in causal discovery is mode (5) as depicted in figure 1: an open, dense, maximal domain of convergence. Achieving mode (5) or any higher mode suffices to provide a unifying treatment for some interesting instances of the nonidentifiability problem in causal discovery.[1]

[1] I would like to thank Kevin Kelly, Jiji Zhang, and Konstantin Genin for the discussion, inspiration, and encouragement they gave me for the current project.

Broader Impact

This work belongs to theoretical computer science and is expected to have contributions to the philosophical discussions about inductive inference.

References

[1] Genin, K. (2018). *The Topology of Statistical Inquiry*. Carnegie Mellon University, PhD Dissertation.

[2] Hunt, Brian R. (1994). "The prevalence of continuous nowhere differentiable functions". In *Proc. Amer. Math. Soc*. American Mathematical Society. 122 (3): 711-717.

[3] Lin, H. (2019). "The hard problem of theory choice: a case study on causal inference and its faithfulness assumption". In *Philosophy of Science*. 86 (5): 967-980.

[4] Lin, H., & Zhang, J. (2020). "On learning causal structures from non-experimental data without any faithfulness assumption". In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, *Proceedings of Machine Learning Research*, 117: 554-582.

[5] Meek, C. (1995) "Strong-completeness and faithfulness in bayesian networks". In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. in:, Montreal, QU, Morgan Kaufmann, San Mateo, CA: 411-418.

[6] Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). "A linear non-Gaussian acyclic model for causal discovery". *Journal of Machine Learning Research*, 7: 2003-2030.

[7] Sprites, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction and search.* Cambridge.

Proofs

Lemma 3 (ϵ-Traction)**.** *Fix a causal graph G on $\{X_1, \ldots, X_n\}$, and let $\mathsf{PA}(X_i)$ be the set of the parents of X_i. Consider any two joint distributions Markov to G and any of their Markov mistures:*

$$\begin{aligned}
P(X_1, \ldots, X_n) &= \prod_{i=1}^{n} P\big(X_i \mid \mathsf{PA}(X_i)\big)\,, \\
Q(X_1, \ldots, X_n) &= \prod_{i=1}^{n} Q\big(X_i \mid \mathsf{PA}(X_i)\big)\,, \\
P_\epsilon(X_1, \ldots, X_n) &= \prod_{i=1}^{n} \Big[(1-\epsilon)P\big(X_i \mid \mathsf{PA}(X_i)\big) \,+\, \epsilon\, Q\big(X_i \mid \mathsf{PA}(X_i)\big)\Big]\,.
\end{aligned}$$

(So $P_0 = P$ and $P_1 = Q$.) Let $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ be a statement of conditional independence about variables $X_1, \ldots, X_n$. Then we have:

1. *If $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ is violated by P or Q, then there exists a (small) real number $\delta > 0$ such that $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ is also violated by P_ϵ for all $\epsilon \in (0, \delta)$.*
2. *P_ϵ is no more than $n\epsilon$ away from P with respect to the total variation metric.*

Proof. Prove clause 1 as follows. Suppose that $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ is violated by P or Q. So that must be violated at some particular assignment of values: $(\mathbf{X} = \mathbf{x}_0, \mathbf{Y} = \mathbf{y}_0, \mathbf{Z} = \mathbf{z}_0)$. Now, use that assignment of values to define a function f as follows, which measures the deviation from conditional independence:

$$\begin{aligned}
f(\epsilon) \;&=_{\text{def}}\; P_\epsilon(\mathbf{X} = \mathbf{x}_0, \mathbf{Y} = \mathbf{y}_0, \mathbf{Z} = \mathbf{z}_0) \cdot P_\epsilon(\mathbf{Z} = \mathbf{z}_0) \\
&\qquad - P_\epsilon(\mathbf{X} = \mathbf{x}_0, \mathbf{Z} = \mathbf{z}_0) \cdot P_\epsilon(\mathbf{Y} = \mathbf{y}_0, \mathbf{Z} = \mathbf{z}_0)\,.
\end{aligned}$$

Since $P_0 = P$ and $P_1 = Q$, the conditional independence is violated by P_0 or P_1 at $(\mathbf{X} = \mathbf{x}_0, \mathbf{Y} = \mathbf{y}_0, \mathbf{Z} = \mathbf{z}_0)$. That is, $f(0) \neq 0$ or $f(1) \neq 0$. It follows that $f(\epsilon)$, as a polynomial in ϵ of order $2n$, is not a zero polynomial and, hence, has at most $2n$ zeros. So $f(\epsilon)$ must be nonzero over a sufficiently short interval of the form $(0, \delta)$. So clause 1 follows. To prove clause 2, define the following joint disatributions, where $k = 1, \ldots, n$:

$$P_\epsilon^{(k)}(X_1, \ldots, X_n) \;=\; \prod_{i=1}^{k} \Big[(1-\epsilon)P(X_i \mid \mathsf{PA}(X_i)) \,+\, \epsilon\, Q(X_i \mid \mathsf{PA}(X_i))\Big] \cdot \prod_{i=k+1}^{n} P(X_i \mid \mathsf{PA}(X_i)).$$

Note that $P_\epsilon^{(k)}$ and $P_\epsilon^{(k+1)}$ differ only in the $(k+1)$-th multiplicative term, with a difference no more than ϵ. Then it is routine to show that $P_\epsilon^{(k)}$ and $P_\epsilon^{(k+1)}$ are no more than ϵ away from each other according to the total variation metric. Then, by the triangular inequality of a metric, $P_\epsilon^{(0)}$ and $P_\epsilon^{(n)}$ are no more than $n\epsilon$ away from each other according to to the total variation metric. But $P_\epsilon^{(0)} = P$ and $P_\epsilon^{(n)} = P_\epsilon$. So clause 2 follows. □

Lemma 4 (Denseness of Faithfulness)**.** *Suppose that $\mathcal{S}$ is inclusive (i.e., closed under both Markov mixtures and graph replacements). Then the faithful Bayes nets in $\mathcal{S}$ form a dense subset of $\mathcal{S}$.*

Proof. Let (G, P) be an arbitrary Bayes net in $\mathcal{S}$. Let $\delta > 0$. It suffices to construct a faithful Bayes net (G, P') in $\mathcal{S}$ whose joint distribution P' is less-then-δ away from P. Enumerate all statements of conditional independence that G does not entail: $\sigma_1, \ldots, \sigma_m$. So, of all the joint distributions Markov to G, there are m ones $Q^{(1)}, \ldots, Q^{(m)}$ that violate $\sigma_1, \ldots, \sigma_m$, respectively. Note that, by closure under graph replacements, the Bayes nets $(G, Q^{(1)}), \ldots, (G, Q^{(m)})$ are all in $\mathcal{S}$. Now, construct a sequence of joint distributions $P^{(1)}, \ldots, P^{(m)}$ as follows: Let ϵ be a nonzero real-valued parameter whose value is to be adjusted later. Let the initial entry of the sequence $P^{(0)} = P$. When $P^{(k-1)}$ has been constructed, construct the next entry of the sequence $P^{(k)}$ as a Markov mixture of the previous entry and $Q^{(k)}$:

$$P^{(k)}(X_1, \ldots, X_n) \;\;=\;\; \prod_{i=1}^{n} \Big[(1-\epsilon)P^{(k-1)}\big(X_i \mid \mathsf{PA}(X_i)\big) \,+\, \epsilon\, Q^{(k)}\big(X_i \mid \mathsf{PA}(X_i)\big)\Big] .$$

Note that, by closure under Markov mixtures and by induction on k, the Bayes nets $(G, P^{(1)}), \ldots, (G, P^{(m)})$ are all in $\mathcal{S}$. Now, find a (sufficiently small) ϵ such that, for any $k = 1, \ldots, m$:

1. $P^{(k)}$ violates σ_k,
2. $P^{(k)}$ also violates all previous statements $\sigma_1, \ldots, \sigma_{k-1}$ (as $P^{(k-1)}$ does),
3. $mn\epsilon < \delta$.

Such an ϵ can be constructed by induction on k and iterated applications of clause 1 of lemma 3. Note that each entry $P^{(k)}$ in the sequence is no more than $n\epsilon$ away from the previous $P^{(k-1)}$ (thanks to clause 2 of lemma 3). So the final entry $P^{(m)}$ is no more than $mn\epsilon$ away from the initial entry $P^{(0)}$ (which is P), by the triangular inequality of a metric. Then, by condition 3 above, P^m is less than δ away from P. It follows that: $(G, P^{(m)})$ is in $\mathcal{S}$ (as established by the two closure properties), faithful (thanks to violation of $\{\sigma_1, \ldots, \sigma_m\}$), and less than δ away from (G, P). This finishes the proof. □

Lemma 5 (Lemma 3 of Lin and Zhang (2020))**.** *Suppose that $(\mathcal{O}, \mathcal{S}, \mathcal{H})$ is a causal learning problem that meets the following conditions:*

A. The state space $\mathcal{S}$ has this closure property: whenever it contains a Bayes net (G, P), then it also contains (G', P) as long as (i) this graph G already occurs in some Bayes net in $\mathcal{S}$ and (ii) this ordered pair (G', P) is indeed a Bayes net.

B. The hypothesis partition $\mathcal{H}$ is this fine-granularity property: it cuts between any faithful causal state and its unfaithful counterparts; that is, for any faithful causal state $s = (G, P) \in \mathcal{S}$ and any of its unfaithful counterparts $s' = (G', P) \in \mathcal{S}$, we have that $H_s \neq H_{s'}$.

Then, with respect to this learning problem, for any learning method M whose domain of convergence is dense and open, we have: M converges to the truth with respect to no unfaithful Bayes net in $\mathcal{S}$ that has a faithful counterpart in $\mathcal{S}$,

We are now ready to prove theorem 1.

Proof of Theorem 1. The preceding lemma 5 applies because condition A follows from the closure of $\mathcal{S}$ under graph replacements and condition B follows from $\mathcal{H}$ being the set of Markov equivalence hypotheses. Then, thanks to lemma 5 and maximality of the domain of convergence, the necessity result follows. It remains to prove the possibility result. Construct a learning method M as follows:

Step 1. Let each conditional independence statement about $\mathcal{O} = \{X_1, \ldots, X_n\}$ be associated with a test that is consistent for any joint distribution that appears in $\mathcal{S}$. Such a test exists on the assumption that $\mathcal{S}$ is regular. Combine those tests into a single "super" test T, which maps each data set D to the set $\Sigma = T(D)$ of all the conditional independence statements accepted by their associated tests given data set D.

Step 2. Linearly order all Markov equivalence hypotheses about $\mathcal{O} = \{X_1, \ldots, X_n\}$ 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 $\mathcal{O}$ to the first hypothesis H_{G_i} in the sequence such that $\mathcal{I}(G_i) \subseteq \Sigma$.

Step 4. Construct learning method $M = F \circ T$.

It is routine to show that M's domain of convergence is maximal and open. It is also dense because it includes the set of all faithful Bayes nets in $\mathcal{S}$, which is a dense subset of $\mathcal{S}$ thanks to lemma 4. □

Proof of Theorem 2. Suppose for reductio that there exists a learning method M for the present learning problem such that its domain of convergence $\mathsf{DC}(M)$ is both dense and open. Consider an arbitrary Bayes net $(\{X \to Y\}, P)$ in $\mathcal{S}$. By denseness, M converges to the truth at some Bayes net $(\{X \to Y\}, P')$ in $\mathcal{S}$. By openness, there exists $\epsilon > 0$ such that M converges to the truth everywhere in the open ϵ-ball in $\mathcal{S}$ centered at $(\{X \to Y\}, P')$. By the closure of $\mathcal{S}$ under graph replacements, the open ϵ-ball in $\mathcal{S}$ centered at $(\{X \to Y\}, P')$ and the open ϵ-ball in $\mathcal{S}$ centered at the revised Bayes net $(\{Y \to X\}, P')$ are isomorphic via the transformation that reverses the causal edge without modifying the joint distribution. It follows that M converges to the truth nowhere in the open ϵ-ball in $\mathcal{S}$ centered at $(\{Y \to X\}, P')$. So the domain of convergence $\mathsf{DC}(M)$ is not dense. Contradiction. □

Proof of Theorem 3. To prove the possibility result, note that it it routine to construct a learning method that converges to the truth at any Bayes net that falls under cases A, B, or C. This domain of convergence is obviously not extendable, so maximality is satisfied. It suffices to show that those cases form a dense and open set of Bayes nets. To see that it is dense, it suffices to observe that the cases in B and those in C already form a dense set. The reason is that, for any Bayes net (G, P) with $X \to Y$ or $Y \to X$, we can find arbitrarily close Bayes nets that violate Gaussianity (in case C); and, for any Bayes net (G, P) with $X \not\to Y$ and $Y \not\to X$, it is automatically in case B. Now, it remains to show that the cases A, B, and C form an open set. To see this, it suffices to note that any sufficiently small change of an instance of case C will still be in case C, any sufficiently small change of an instance of case B will still be in case B, and any sufficiently small change of an instance of case$A$ will still be in case A or case C. This completes the proof.

Now, establish clause 2 contrapositively: Suppose that a learning method M fails to converge to the truth in a non-degenerate non-Gaussian Bayes net (G, P). It suffices to show that its domain of convergence is not maximal. First, construct a consistent test T of the hypothesis that the true distribution is identical to P. Second, construct another learning method M' that outputs the causal hypothesis true of G when T says "yes" and otherwise outputs whatever M outputs. And finally, show that M' has a more inclusive domain of convergence: $\mathsf{DC}(M') = \mathsf{DC}(M') \cup \{(G, P)\}$, which implies that M's domain of convergence is not maximal, as desired.

To establish clause 3, it suffices to note that the convergence property has to be sacrificed for case B', by lemma 5, and hence it has to be secured for case B by maximality. □

Proof of Lemma 2. Suppose for reductio that there exists learning method M such that its domain of convergence $\mathsf{DC}(M)$ is maximal, dense, but *not* open. Since $\mathsf{DC}(M)$ is not open, there exists a Bayes net $s = (G, P) \in \mathsf{DC}(M)$, but we can always find a Bayes nets $s' = (G, P')$ arbitrarily close to s such that $s' \notin \mathsf{DC}(M)$. Let ϵ be a positive real-valued parameter, to be fine-tuned later. Since $s \in \mathsf{DC}(M)$ Let n_0 be a sample size at which M outputs the truth in s with a probability at least $1-\epsilon$. Since we can always find a Bayes nets $s' = (G, P')$ arbitrarily close to s such that $s' \notin \mathsf{DC}(M)$, let P' be so close to P that, at any sample size no more than n_1, the sampling distribution in s' and the sampling distribution in s are at most ϵ away from each other (according to the total variation metric). It follows that, at sample size n_1, M outputs the truth at $s' = (G, P')$ with a probability at least $1 - 2\epsilon$. Now, since $s' = (G, P') \notin \mathsf{DC}(M)$, there exists another Bayes net $s'' = (G', P') \in \mathcal{S}$ such that $s'' \in \mathsf{DC}(M)$, for otherwise we would be able to construct another learning method M' with a more inclusive domain of convergence $\mathsf{DC}(M') = \mathsf{DC}(M) \cup \{s'\}$, which contradicts the maximality of $\mathsf{DC}(M)$. Since $s'' \in \mathsf{DC}(M)$, there exists a sample size $n_2 > n_1$ at which M outputs the truth $H_{s''}$ in $s'' = (G', P')$ with a probability at least $1 - 2\epsilon$. Then, since s' and s'' share the same joint distribution, it follows that, at sample size n_2, M outputs the falsehood $H_{s''}$ in $s' = (G, P')$ with a probability at least $1 - 2\epsilon$, and hence outputs the truth in the same Bayes net $s' = (G, P')$ with a probability at most 2ϵ. Here is an interim summary: in $s' = (G, P')$, the probability for M to output the truth goes up to at least $1 - 2\epsilon$ at sample size n_1 and then drops and becomes at most 2ϵ at sample size n_2. No, since $\mathsf{DC}(M)$ is dense, we can always find a Bayes net $s''' = (G, P''')$ arbitrarily close to $s' = (G, P')$ such that $s''' \in \mathsf{DC}(M)$. Then let P''' be so close that P' that, at any sample size no more than n_2, the sampling distribution in s' and the sampling distribution in s are at most ϵ away from each other (according to the total variation metric). So in $s''' = (G, P''')$, the probability for M to output the truth goes up to at least $1 - 3\epsilon$ at sample size n_1 and then drops and becomes at most 3ϵ at sample size n_2. So M fails to converge the truth $(1 - 6\epsilon)$-progressively. Now, adjust the value of parameter ϵ to ensure that $1 - 6\epsilon = \alpha$. This finishes the proof. □