The Causal-Neural Connection: Expressiveness, Learnability, and Inference

Anonymous Author(s)
Affiliation
Address
email

Abstract

One of the central elements of any causal inference is a structural causal model (SCM), which represents a collection of mechanisms and exogenous sources of random variation of the system under investigation (Pearl, 2000). An important property of many kinds of neural networks is *universal approximability*: the ability to approximate any function to arbitrary precision. Given this property, one may be tempted to surmise that a collection of neural nets is capable of learning any SCM by training on data generated by that SCM. In this paper, we show this is not the case by disentangling the notions of expressivity and learnability. Specifically, we show that the causal hierarchy theorem (Thm. 1, Bareinboim et al., 2020), which describes the limits of what can be learned from data, still holds for neural models. For instance, an arbitrarily complex and expressive neural net is unable to predict the effects of interventions given observational data alone. Given this result, we introduce a special type of SCM called a neural causal model (NCM), and establish additional inductive biases in the form of structural constraints for performing causal inferences within this class of models; we focus on the tasks known as causal identification and estimation. Leveraging the neural toolbox, we develop an algorithm that is both sufficient and necessary to determine whether a causal effect can be learned from data (i.e., causal identifiability); if identifiability holds, it also estimates the effect from data (causal estimation). Simulations corroborate the proposed approach.

1 Introduction

One of the most celebrated and relied upon results in the science of intelligence is the universality of neural models. More formally, universality says that neural models can approximate any function (e.g., boolean, classification boundaries, continuous valued) with arbitrary precision given enough capacity in terms of the depth and breadth of the network [10, 21, 39, 45]. This result, combined with the observation that most tasks can be abstracted away and modeled as input/output – i.e., as functions – lead to the strongly held belief that under the right conditions, neural networks can solve the most challenging and interesting tasks in AI. This belief is not without merits, and is corroborated by ample evidence of practical successes, including in compelling tasks in computer vision [36], speech recognition [17], and game playing [46]. Given that the universality of neural nets is such an appealing proposition, we investigate through causal lenses whether this belief is well-founded.

To start understanding the causal-neural connection – i.e., the non-trivial and somewhat intricate relationship between these modes of reasoning – two standard objects in causal analysis will be instrumental. First, we evoke a class of generative models known as the *Structural Causal Model* (SCM, for short) [48, Ch. 7]. In words, an SCM $\mathcal{M}^*$ is a representation of a system that includes a collection of mechanisms and a probability distribution over the exogenous conditions (to be formally defined later on). Second, any fully specified SCM $\mathcal{M}^*$ induces a collection of distributions known as the *Pearl Causal Hierarchy* (PCH) [5, Def. 9]. The importance of the PCH is that it formally delimits

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

distinct cognitive capabilities (also known as layers) that can be associated with the human activities of "seeing" (layer 1), "doing" (2), and "imagining" (3) [49, Ch. 1]. [1] Each of these layers can be expressed as a distinct formal language and represents queries that can help to classify different types of inferences. Together, these layers form a strict containment hierarchy [5, Thm. 1]. We illustrate these notions in Fig. 1(a) (left side), where SCM $\mathcal{M}^*$ induces layers L_1^*, L_2^*, L_3^* of the PCH.

Even though each possible statement within these capabilities has well-defined semantics given the true SCM $\mathcal{M}^*$ [48, Ch. 7], a challenging inferential task arises when one wishes to recover part of $\mathcal{M}^*$ when it is only partially observed. This situation is typical in the real world, aside from some special settings in physics and chemistry where the laws of nature are understood with high precision. It can lead to challenges found in many practical settings – for example, when one needs to make a statement about the effect of a new intervention (i.e., about layer 2), but only has observational data from layer 1, which is passively collected.[2]

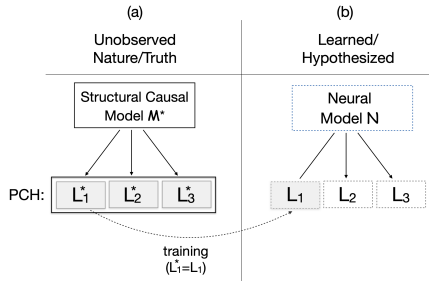

Figure 1: The l.h.s. contains unobserved SCM $\mathcal{M}^*$ that induces the three layers of the PCH. The r.h.s. contains NCM that is trained and matches layer 1. The match in the shade represents that the two models agree with respect to L_1 while not necessarily agree w.r.t layers 2 and 3.

Going back to the causal-neural connection to try to address this inferential challenge, one could consider the following idea – learn a neural model $\mathcal{N}$ using the observational dataset (layer 1) generated by the true SCM $\mathcal{M}^*$, as illustrated in Fig. 1(b). Naturally, a basic consistency requirement would be that $\mathcal{N}$ should be capable of generating the same distribution as the one coming from $\mathcal{M}^*$; in this case, their layer 1 predictions should match (i.e., $L_1 = L_1^*$). Given the universality of neural models, it is not hard to believe that this can be accomplished in the large sample limit. The question arises on whether the learned model $\mathcal{N}$ can act as a proxy, having the capability of predicting the effect of interventions that matches the L_2-distribution generated by the true (unobserved) SCM $\mathcal{M}^*$. [3] This cannot be ascertained in general, as will become evident later on (Cor. 1). The intuitive reason behind this result is that there are multiple neural models that are equally consistent w.r.t. the L_1 distribution but generate different L_2-distributions. [4] By and large, even though $\mathcal{N}$ may be expressive enough to fully represent $\mathcal{M}^*$ (as discussed later on), generating one particular parametrization of $\mathcal{N}$ (consistent with L_1) is insufficient to provide any guarantee regarding higher-layer inferences (e.g., L_2, L_3).

The discussion above entails two tasks that have been acknowledged in the literature, namely, causal effect identification and estimation. The first – causal identification – has been extensively studied, and general solutions have been developed such as Pearl's celebrated do-calculus [47]. Given the impossibility described above, the ingredient shared across current non-neural solutions is to represent assumptions about the unknown $\mathcal{M}^*$ in the form of causal diagrams [48, 55, 4] or their equivalence classes [23, 50, 24, 60]. There are no neural methods today focused on solving such a task.

The second task – causal estimation – is triggered when effects are identifiable by the first task. Estimating effects when the backdoor/conditional ignorability condition holds [48, Sec. 3.3.1] has attracted interest in the deep learning community, and impressive results have been achieved [53, 44, 40, 25, 58, 59, 30, 54, 11, 20]. There are also non-neural methods for the estimation of identifiable functionals that are not of the backdoor-form [26, 27, 28]. Each kind of approach

[1] This structure is named after Judea Pearl and is a central topic in his *Book of Why (BoW)*, where it is also called the "Ladder of Causation" [49]. For further discussion, see [5].

[2] This inferential challenge is more general since an agent may have interventional data from any subset of the layers, while aiming to make inferences about some other parts of the PCH [4, 5]; this is quite common in RL settings [56, 12, 38, 37]. Still, for the sake of space and concreteness, we will focus on a specific instance.

[3] We defer a more formal discussion on how neural models could be used to assess the effect of interventions to Sec. 2. Still, this is neither attainable in all universal neural architectures nor obvious how to be implemented.

[4] Pearl shared a similar observation in the *BoW* [49, p. 32]: "Without the causal model, we could not go from rung (layer) one to rung (layer) two. This is why deep-learning systems (as long as they use only rung-one data and do not have a causal model) will never be able to answer questions about interventions (...)".

optimizes an estimand corresponding to one specific target effect; neither provides a generative model that can act as a proxy, with the power of computing any identifiable effect as if it were the true $\mathcal{M}^*$.

Despite all the great progress achieved so far, it is still largely unknown how to perform the tasks of causal identification and estimation in general settings using neural models as a generative model, acting as a proxy of the true SCM $\mathcal{M}^*$. It is our goal here to develop a general causal-neural framework that has the potential to scale to real-world, high-dimensional data domains while preserving the validity of the inferences like the traditional, symbolic approaches. In the same way that the causal diagram encodes the assumptions necessary for the do-calculus to decide whether a certain query is identifiable, our method uses an equivalent inductive bias while being amenable to gradient-based optimization, allowing us to perform both tasks in an integrated fashion (in a way, realizing Pearl's observation raised in Footnote 4). More specifically, our contributions are as follows:

1. [Sec. 2] We introduce a simple yet special type of SCM that is more amenable to neural computation called a *neural causal model* (NCM). We prove basic properties of this class of models, including its universal expressiveness and capability of encoding causal invariances (Thm. 1-3). Notably, we show that despite the NCM's expressivity, it is still constrained by the Causal Hierarchy Theorem, which means it cannot perform causal inferences without causal assumptions (Corol. 1).

2. [Sec. 3] We formalize the problem of neural identification (Def. 8) and prove a duality between identification in causal diagrams and in neural causal models (Thm. 4). We introduce an operational way to perform inferences in NCMs (Corol. 2-3) and a sound and complete algorithm to jointly train and decide effect identifiability for an NCM (Alg. 1, Corol. 4)

3. [Sec. 4] Building on these results, we develop an algorithm to jointly identify and estimate causal effects amenable to gradient-based learning (Alg. 2). [Sec. 5] Simulations corroborate our method.

1.1 Preliminaries

In this section, we provide the necessary background to understand this work following the presentation in [48]. An uppercase letter X indicates a random variable, and a lowercase letter x indicates its corresponding value; bold uppercase $\mathbf{X}$ denotes a set of random variables, and lowercase letter $\mathbf{x}$ its corresponding values. We use $\mathcal{D}_X$ to denote the domain of X and $\mathcal{D}_{\mathbf{X}} = \mathcal{D}_{X_1} \times \cdots \times \mathcal{D}_{X_k}$ for $\mathbf{X} = \{X_1, \ldots, X_k\}$. We denote $P(\mathbf{X})$ as a probability distribution over a set of random variables $\mathbf{X}$ and $P(\mathbf{X} = \mathbf{x})$ as the probability of $\mathbf{X}$ being equal to the value of $\mathbf{x}$ under the distribution $P(\mathbf{X})$. For simplicity, we will mostly abbreviate $P(\mathbf{X} = \mathbf{x})$ as simply $P(\mathbf{x})$. The basic semantic framework of our analysis rests on *structural causal models* (SCMs) [48, Ch. 7], which are defined below.

Definition 1 (Structural Causal Model (SCM))**.** An SCM $\mathcal{M}$ is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$, where $\mathbf{U}$ is a set of exogenous variables (or "error terms") that are determined by factors outside the model; $\mathbf{V}$ is a set $\{V_1, V_2, \ldots, V_n\}$ of (endogenous) variables of interest that are determined by other variables in the model – that is, in $\mathbf{U} \cup \mathbf{V}$; $\mathcal{F}$ is a set of functions $\{f_{V_1}, f_{V_2}, \ldots, f_{V_n}\}$ such that each f_i is a mapping from (the respective domains of) $\mathbf{U}_{V_i} \cup \mathbf{Pa}_{V_i}$ to V_i, where $\mathbf{U}_{V_i} \subseteq \mathbf{U}$, $\mathbf{Pa}_{V_i} \subseteq \mathbf{V} \setminus V_i$, and the entire set $\mathcal{F}$ forms a mapping from $\mathbf{U}$ to $\mathbf{V}$. That is, for $i = 1, \ldots, n$, each $f_i \in \mathcal{F}$ is such that $v_i \leftarrow f_{V_i}(\mathbf{pa}_{V_i}, \mathbf{u}_{V_i})$; and $P(\mathbf{u})$ is a probability function defined over the domain of $\mathbf{U}$. ■

Each SCM $\mathcal{M}$ induces a causal diagram G where every $V_i \in \mathbf{V}$ is a vertex, there is a directed arrow $(V_j \rightarrow V_i)$ for every $V_i \in \mathbf{V}$ and $V_j \in Pa(V_i)$, and a dashed-bidirected arrow $(V_j \dashleftarrow\dashrightarrow V_i)$ for every pair $V_i, V_j \in \mathbf{V}$ such that $\mathbf{U}_i \cap \mathbf{U}_j \neq \emptyset$. For further details on this construction, see [5, Def. 13/16, Thm. 4]. The exogenous $\mathbf{U}_i$'s are not assumed independent (i.e. Markovianity does not hold). Our treatment is constrained to recursive SCMs, which implies acyclic causal diagrams.

We show next how an SCM $\mathcal{M}$ gives values to the PCH's layers (not to be confused with neural net layers); for details on the semantics, see [5, Sec. 1.2]. Superscripts are omitted when unambiguous.

Definition 2 (Layers 1, 2 Valuations)**.** An SCM $\mathcal{M}$ induces layer $L_2(\mathcal{M})$, a set of distributions over $\mathbf{V}$, one for each intervention $\mathbf{x}$. For each $\mathbf{Y} \subseteq \mathbf{V}$,

$$P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}) = \sum_{\{\mathbf{u} | \mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}\}} P(\mathbf{u}), \tag{1}$$

where $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ is the solution for $\mathbf{Y}$ after evaluating $\mathcal{F}_{\mathbf{x}} := \{f_{V_i} : V_i \in \mathbf{V} \setminus \mathbf{X}\} \cup \{f_X \leftarrow x : X \in \mathbf{X}\}$. The specific distribution $P(\mathbf{V})$, where $\mathbf{X}$ is empty, is defined as layer $L_1(\mathcal{M})$. ■

In words, external intervention forcing a set of variables $\mathbf{X}$ to take values $\mathbf{x}$ are modeled by replacing the original mechanism f_X for each $X \in \mathbf{X}$ with its corresponding value in $\mathbf{x}$. This operation is known as the do-operator, $do(\mathbf{X} = \mathbf{x})$ and graphically as the *mutilation* procedure.

2 Neural Causal Models and the Causal Hierarchy Theorem

In this section, we discuss the tension between expressiveness and learnability (Fig. 1(a)). To that end, we define a special class of SCMs based on neural nets that is amenable to optimization. This data structure should have the potential to act as a proxy of the true, unobserved SCM $\mathcal{M}^*$ (Fig. 1(b)).

Definition 3 (NCM)**.** A Neural Causal Model (for short, NCM) $\widehat{M}(\boldsymbol{\theta})$ over variables $\mathbf{V}$ with parameters $\boldsymbol{\theta} = \{\theta_{V_i} : V_i \in \mathbf{V}\}$ is an SCM $\langle \widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, \widehat{P}(\widehat{\mathbf{U}})\rangle$ such that

- $\widehat{\mathbf{U}} \subseteq \{\widehat{U}_{\mathbf{C}} : \mathbf{C} \subseteq \mathbf{V}\}$, where each $\widehat{U}$ is associated with some subset of variables $\mathbf{C} \subseteq \mathbf{V}$, and $\mathcal{D}_{\widehat{U}} = [0, 1]$ for all $\widehat{U} \in \widehat{\mathbf{U}}$. (Unobserved confounding is present whenever $|\mathbf{C}| > 1$.)
- $\widehat{\mathcal{F}} = \{\hat{f}_{V_i} : V_i \in \mathbf{V}\}$, where each $\hat{f}_{V_i}$ is a feedforward neural network parameterized by $\theta_{V_i} \in \boldsymbol{\theta}$ mapping values of $\mathbf{U}_{V_i} \cup \mathbf{Pa}_{V_i}$ to values of V_i for some $\mathbf{Pa}_{V_i} \subseteq \mathbf{V}$ and $\mathbf{U}_{V_i} = \{\widehat{U}_{\mathbf{C}} : \widehat{U}_{\mathbf{C}} \in \widehat{\mathbf{U}}, V_i \in \mathbf{C}\}$;
- $\widehat{P}(\widehat{\mathbf{U}})$ is defined s.t. $\widehat{U} \sim \text{Unif}(0, 1)$ for each $\widehat{U} \in \widehat{\mathbf{U}}$. ■

Some remarks are worth making at this point. First, by definition, all NCMs are SCMs, so they have the capability of generating any distributions associated to the PCH's layers. Second, not all SCMs are NCMs since Def. 3 dictates that $\widehat{\mathbf{U}}$ follows uniform distributions in the unit interval and $\widehat{\mathcal{F}}$ are feedforward neural networks. [5] Also, note that each exogenous variable $\mathbf{U}_{V_i} \in \widehat{\mathbf{U}}$ is not independent, which will play a critical role in causality, allowing for the existence of unobserved confounding. To compare the expressiveness of NCMs and SCMs, we formalize the notion of consistency.

Definition 4 ($\text{P}^{(L_i)}$-Consistency)**.** Consider two SCMs, $\mathcal{M}_1$ and $\mathcal{M}_2$. $\mathcal{M}_2$ is said to be $\text{P}^{(L_i)}$-consistent (for short, L_i-consistent) w.r.t. $\mathcal{M}_1$ if $L_i(\mathcal{M}_1) = L_i(\mathcal{M}_2)$. ■

This definition applies to NCMs since they are also SCMs. As shown below, NCMs can not only approximate the collection of functions of the true SCM $\mathcal{M}^*$, but they can perfectly *represent* all the observational, interventional, and counterfactual distributions. This property is, in fact, special and not enjoyed by many neural models. (For examples and a more detailed discussion, see Appendix C.)

Theorem 1 (NCM Expressiveness)**.** *For any SCM $\mathcal{M}^* = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U})\rangle$, there exists an NCM $\widehat{M}(\boldsymbol{\theta}) = \langle \widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, \widehat{P}(\widehat{\mathbf{U}})\rangle$ s.t. $\widehat{M}$ is L_3-consistent w.r.t. $\mathcal{M}^*$.* ■

Due to space constraints, proofs are provided in Appendix A. Thm. 1 ascertains that there is no loss of expressive power using NCMs despite the constraints imposed over its form, i.e., NCMs are as expressive as SCMs. One might be tempted to surmise, therefore, that an NCM can be trained on the observed data and act as a proxy for the true SCM $\mathcal{M}^*$, i.e., inferences about other quantities of $\mathcal{M}^*$ can be done through computation directly in $\widehat{\mathcal{M}}$. Unfortunately, this is almost never the case: [6]

Corollary 1 (Neural Causal Hierarchy Theorem (N-CHT))**.** *Let Ω^* and Ω be the sets of all SCMs and NCMs, respectively. We say that Layer j of the causal hierarchy for NCMs* collapses *to Layer i ($i < j$) relative to $\mathcal{M}^* \in \Omega^*$ if $L_i(\mathcal{M}^*) = L_i(\widehat{M})$ implies that $L_j(\mathcal{M}^*) = L_j(\widehat{M})$ for all $\widehat{M} \in \Omega$. Then, with respect to the Lebesgue measure over (a suitable encoding of L_3-equivalence classes of) SCMs, the subset in which Layer j of NCMs collapses to Layer i has measure zero.* ■

This corollary highlights the fundamental challenge of performing inferences across the PCH layers even when the target object (NCM $\widehat{\mathcal{M}}$) is a suitable surrogate for the underlying SCM $\mathcal{M}^*$, in terms of expressiveness and capability of generating the same observed distribution. I.e., expressiveness does not mean that the learned object has the same empirical content as the generating model.

2.1 A Family of Neural-Interventional Constraints

In this section, we investigate constraints about $\mathcal{M}^*$ that will narrow down the hypothesis space and possibly allow for these cross-layer inferences. One well-studied family of structural constraints comes in the form of a pair comprised of a causal diagram $\mathcal{G}$ and collection of interventional distributions $\mathcal{P}$, known as a *causal bayesian network* (CBN) [5,

[5] The implications of these choices will become apparent later on, but intuitively, feedforward networks are universal approximators [10, 21] (see also [14]), and any other probability distribution can be generated by the uniform distribution (e.g., see *probability integral transform* [1]). This suggests that the pair $\langle \widehat{\mathcal{F}}, \widehat{P}(\widehat{\mathbf{U}})\rangle$ may be a suitable choice, and expressive enough for modeling $\mathcal{M}^*$'s mechanisms $\mathcal{F}$ and distribution $P(\mathbf{U})$.

[6] Multiple examples of this phenomenon are discussed in the Appendix C.1.

Def. 16; Thm. 4]. The diagram $\mathcal{G}$ encodes constraints over the space of interventional distributions $\mathcal{P}$ which are useful to perform cross-layer inferences (for details, see Appendix A). For simplicity, we focus on interventional inferences from observational data. To compare the constraints entailed by distinct SCMs, we define the following notion of consistency:

Definition 5 ($\mathcal{G}$-Consistency)**.** Let $\mathcal{G}$ be the causal diagram induced by SCM $\mathcal{M}^*$. For any SCM $\mathcal{M}$, we say that $\mathcal{M}$ is $\mathcal{G}$-consistent (w.r.t. $\mathcal{M}^*$) if $\mathcal{G}$ is a CBN for $L_2(\mathcal{M})$. ■

In the context of NCMs, this means that $\mathcal{M}$ would impose the same constraints over $\mathcal{P}$ as the true SCM $\mathcal{M}^*$. Whenever the corresponding diagram $\mathcal{G}$ is known [7], one should only consider NCMs that are $\mathcal{G}$-consistent. We provide below a systematic way of constructing $\mathcal{G}$-consistent NCMs.

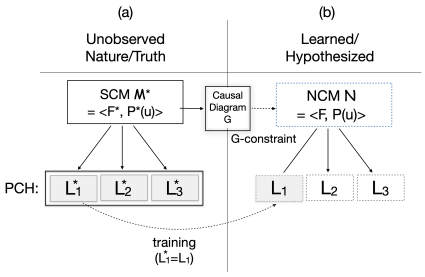

Figure 2: The l.h.s. contains the true SCM $\mathcal{M}^*$ that induces PCH's three layers. The r.h.s. contains an NCM that is trained with layer 1 data. The match in the shade represents that the two models agree with respect to L_1 while not necessarily agree in layers 2 and 3.

Definition 6 (C^2-Component)**.** For a causal diagram $\mathcal{G}$, a subset $\mathbf{C} \subseteq \mathbf{V}$ is a complete confounded component (for short, C^2-component) if any pair $V_i, V_j \in \mathbf{C}$ is connected with a bidirected arrow in $\mathcal{G}$ and is maximal (i.e. there is no C^2-component $\mathbf{C}'$ for which $\mathbf{C} \subset \mathbf{C}'$.) ■

Definition 7 ($\mathcal{G}$-Constrained NCM (constructive))**.** Let $\mathcal{G}$ be the causal diagram induced by SCM $\mathcal{M}^*$. Construct NCM $\widehat{M}$ as follows. **(1)** Choose $\widehat{\mathbf{U}}$ s.t. $\widehat{U}_{\mathbf{C}} \in \widehat{\mathbf{U}}$ if and only if $\mathbf{C}$ is a C^2-component in $\mathcal{G}$. **(2)** For each variable $V_i \in \mathbf{V}$, choose $\mathbf{Pa}_{V_i} \subseteq \mathbf{V}$ s.t. for every $V_j \in \mathbf{V}$, $V_j \in \mathbf{Pa}_{V_i}$ if and only if there is a directed edge from V_j to V_i in $\mathcal{G}$. Any NCM in this family is said to be $\mathcal{G}$-constrained. ■

Note that this represents a family of NCMs, not a unique one, since $\boldsymbol{\theta}$ (the parameters of the neural networks) are not yet specified by the construction, only the scope of the function and independence relations among the sources of randomness ($\widehat{\mathbf{U}}$). In contrast to the original SCM where both $\langle \mathcal{F}, P(\mathbf{u}) \rangle$ can freely vary, the degrees of freedom within NCMs come from $\boldsymbol{\theta}$. [8]

We show next that an NCM constructed following the procedure dictated by Def. 7 encodes all the constraints of the original causal diagram.

Theorem 2 (NCM $\mathcal{G}$-Consistency)**.** *Let $\widehat{M}(\boldsymbol{\theta})$ be a $\mathcal{G}$-constrained NCM. $\widehat{\mathcal{M}}$ is $\mathcal{G}$-consistent.* ■

We show next the implications of imposing the structural constraints embedded in the causal diagram.

Theorem 3 (L_2-$\mathcal{G}$ Representation)**.** *For any SCM $\mathcal{M}^*$ that induces causal diagram $\mathcal{G}$, there exists a $\mathcal{G}$-constrained NCM $\widehat{M}(\boldsymbol{\theta}) = \langle \widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, \widehat{P}(\widehat{\mathbf{U}}) \rangle$ that is L_2-consistent w.r.t. $\mathcal{M}^*$.* ■

The importance of this result stems from the fact that despite constraining the space of NCMs for the ones compatible with $\mathcal{G}$, the resultant family is still expressive enough to represent the entire Layer 2 of the original, unobserved SCM $\mathcal{M}^*$.

Fig. 2 provides a mental picture useful to understand the results discussed in this section. The true SCM $\mathcal{M}^*$ generates the three layers of the causal hierarchy (left side), but in many settings only observational data (layer 1) is visible. An NCM $\widehat{M}$ trained with this data is capable of perfectly representing L_1 (right side). For almost any generating $\mathcal{M}^*$ sampled from Ω^*, there exists an NCM $\widehat{M}$ that exhibits the same behavior with respect to the observational data ($\widehat{M}$ is L_1-consistent) but exhibits a different behavior with respect to the interventional one. In other words, L_1 underdetermines L_2. (Similarly, L_1 and L_2 underdetermine L_3.) Still, if we impose G-consistency in the construction of the NCM, $\widehat{M}$ may agree with the true $\mathcal{M}^*$ under some general conditions, but not always, which we'll investigate in the next section.

[7] Otherwise, the causal diagram can be learned through structural learning algorithms from observational data [55, 51] or experimental data [35, 34, 22]. See the next footnote for a neural approach for this task.

[8] There are some works in the literature that model SCMs using neural networks as functions, but which differ in nature and scope to our work. [16] attempts to learn the entire SCM from observational data under the Markov assumptions. This entails strong constraints over $P(U)$, which in the context of identification means all effects are always identifiable; see Corol. 3. [7, 9] also use experimental (L_2) data to learn the causal diagram $\mathcal{G}$.

3 The Neural Identification Problem

We now investigate the feasibility of causal inferences in the class of $\mathcal{G}$-constrained NCMs. [9] The first step is to refine the notion of identification [48, pp. 67] to inferences within this class of models.

Definition 8 (Neural Effect Identification)**.** Let Ω^* be the set of all SCMs, and Ω the set of NCMs. The causal effect $P(\mathbf{y} \mid do(\mathbf{x}))$ is said to be neural identifiable from the set of $\mathcal{G}$-constrained NCMs, $\Omega(\mathcal{G})$, and observational data $P(\mathbf{v})$ if and only if $P^{\mathcal{M}^*}(\mathbf{y} \mid do(\mathbf{x})) = P^{\widehat{M}}(\mathbf{y} \mid do(\mathbf{x}))$ for every pair of models $\mathcal{M}^* \in \Omega^*$ and $\widehat{M} \in \Omega(\mathcal{G})$ s.t. $\mathcal{M}^*$ induces $\mathcal{G}$ and $P(\mathbf{v}) = P^{\mathcal{M}^*}(\mathbf{v}) = P^{\widehat{M}}(\mathbf{v})$. ■

In the context of graphical identifiability [48, Def. 3.2.4] and the do-calculus, an effect is identifiable if any SCM in Ω^* compatible with the observed causal diagram and capable of generating the observational distribution match the interventional query. If we constrained attention to NCMs, identification in the general class would imply identification in NCMs, naturally, since it needs to hold for all SCMs. On the other hand, it would be insufficient to constrain identification within the NCM class since it is conceivable that the effect could match within the class (perhaps in a not very expressive neural architecture) while there still exists an SCM that generates the same observational distribution, induce the same graph, but does not agree in the interventional distribution. (See Example 2 in Appendix C.) Accordingly, Def. 8 relates the solution space of these two classes of models and require the solution within the neural class to match the solution within the SCM class; for an illustration, see Fig. 3. The next result makes the relationship between these classes more explicit.

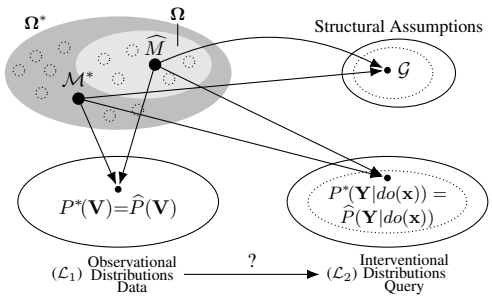

Figure 3: $P(\mathbf{Y} \mid do(\mathbf{x}))$ is identifiable from $P(\mathbf{V})$ and NCM $\widehat{\mathcal{M}} \in \Omega$ if for any SCM $\mathcal{M}^* \in \Omega^*$ (top left), $\widehat{\mathcal{M}}, \mathcal{M}^*$ match in $P(\mathbf{V})$ (bottom left) and $\mathcal{G}$ (top right), they also match in $P(\mathbf{Y} \mid do(\mathbf{x}))$ (bottom right).

Theorem 4 (Graphical-Neural Equivalence (Dual ID))**.** *Let Ω^* be the set of all SCMs and Ω the set of NCMs. Consider the true SCM $\mathcal{M}^*$ and the corresponding causal diagram $\mathcal{G}$. Let $Q = P(\mathbf{y} \mid do(\mathbf{x}))$ be the query of interest and $P(\mathbf{v})$ the observational distribution. Then, Q is identifiable from $\Omega(\mathcal{G})$ and $P(\mathbf{v})$ if and only if it is identifiable from $\mathcal{G}$ and $P(\mathbf{v})$.* ■

Theorem 4 says that the identification status of a query is preserved across settings. For instance, if an effect is identifiable from the combination of a causal graph $\mathcal{G}$ and $P(\mathbf{v})$, it will also be identifiable from $\mathcal{G}$-constrained NCM (and the other way around). This is encouraging since our goal is to perform inferences directly through neural causal models, avoiding the symbolic nature of do-calculus computation; the theorem guarantees that this is, in principle, achievable.

Corollary 2 (Neural Mutilation (Operational ID))**.** *Consider the true SCM $\mathcal{M}^* \in \Omega^*$, causal diagram $\mathcal{G}$, the observational distribution $P(\mathbf{v})$, and a target query Q equal to $P^{\mathcal{M}^*}(\mathbf{y} \mid do(\mathbf{x}))$. Let $\widehat{\mathcal{M}} \in \Omega(\mathcal{G})$ be a $\mathcal{G}$-constrained NCM that is L_1-consistent with $\mathcal{M}^*$. If the effect is identifiable from $\mathcal{G}$ and $P(\mathbf{v})$, then Q is computable through a mutilation process on a proxy NCM $\widehat{\mathcal{M}}$, i.e., for each $X \in \mathbf{X}$, replacing the equation f_x with a constant x ($Q = \text{PROC-MUTILATION}(\widehat{M}; \mathbf{X}, \mathbf{Y})$).* ■

Following the duality stated by Thm. 4, this result provides a practical, operational way of evaluating queries in NCMs: inferences may be carried out through the process of mutilation, which gives semantics to queries in the generating SCM $\mathcal{M}^*$ (via Def. 2). What is interesting here is that the proposition provides conditions under which this process lead to valid inferences, even when $\mathcal{M}^*$ is unknown or when the mechanisms $\mathcal{F}$ and unobserved noise $P(\mathbf{U})$ of $\mathcal{M}^*$ and proxy NCM $\widehat{M}$ do not match. Inferences using mutilation on $\widehat{M}$ would work as if they were on $\mathcal{M}^*$ itself, and they would be correct so long as certain stringent properties were satisfied – L_1-consistency, G-constraint, and identifiability. As shown earlier, if these properties are not satisfied, inferences within a proxy model will almost never be valid, bearing no relationship with the ground truth.

[9] This is akin to what happens with the non-neural CHT and the subsequent use of causal diagrams to encode the necessary inductive bias, and in which the do-calculus allows for cross-layer inferences directly from the graphical representation [5, Sec. 1.4].

Still, one special class of SCMs in which any interventional distribution is identifiable is called Markovian, where all U_i are assumed independent and affect only one endogenous variable V_i.

Corollary 3 (Markovian Identification)**.** *Whenever the $\mathcal{G}$-constrained NCM $\widehat{\mathcal{M}}$ is Markovian, $P(\mathbf{y} \mid do(\mathbf{x}))$ is always identifiable through the process of mutilation in the proxy NCM (via Corol. 2).* ■

This is obviously not the case for general non-Markovian models, which leads to the very problem of identification. In these cases, we need to decide whether the mutilation procedure (Corol. 2) will, in principle, give the correct answer. We show in Alg. 1 a learning procedure that is capable of deciding whether a certain effect is identifiable from observational data. Remarkably, the procedure is both necessary and sufficient, which means that all, and only, identifiable effects are classified as such by our procedure. This implies that, in principle, deep learning could be as powerful as the do-calculus to decide identifiability. (For a detailed discussion of non-identifiability and examples, see Appendix C.3.)

Algorithm 1: Identifying/estimating queries with NCMs.

Input : causal query $Q = P(\mathbf{y} \mid do(\mathbf{x}))$, L_1 data $P(\mathbf{v})$, and causal diagram $\mathcal{G}$

Output : $P^{\mathcal{M}^*}(\mathbf{y} \mid do(\mathbf{x}))$ if identifiable, `FAIL` otherwise.

$\widehat{M} \leftarrow$ `NCM`$(\mathbf{V}, \mathcal{G})$ `// from Def. 7`

$\boldsymbol{\theta}^*_{\min} \leftarrow \arg\min_{\boldsymbol{\theta}} P^{\widehat{M}(\boldsymbol{\theta})}(\mathbf{y} \mid do(\mathbf{x}))$ s.t. $L_1(\widehat{M}(\boldsymbol{\theta})) = P(\mathbf{v})$

$\boldsymbol{\theta}^*_{\max} \leftarrow \arg\max_{\boldsymbol{\theta}} P^{\widehat{M}(\boldsymbol{\theta})}(\mathbf{y} \mid do(\mathbf{x}))$ s.t. $L_1(\widehat{M}(\boldsymbol{\theta})) = P(\mathbf{v})$

if $P^{\widehat{M}(\boldsymbol{\theta}^*_{\min})}(\mathbf{y} \mid do(\mathbf{x})) \neq P^{\widehat{M}(\boldsymbol{\theta}^*_{\max})}(\mathbf{y} \mid do(\mathbf{x}))$ **then**

  return `FAIL`

else

  return $P^{\widehat{M}(\boldsymbol{\theta}^*_{\min})}(\mathbf{y} \mid do(\mathbf{x}))$ `// choose min or max arbitrarily`

Corollary 4 (Soundness and Completeness)**.** *Let Ω^* be the set of all SCMs, $\mathcal{M}^* \in \Omega^*$ be the true SCM inducing causal diagram $\mathcal{G}$, and $Q = P(\mathbf{y} \mid do(\mathbf{x}))$ be a query of interest. Assume $\widehat{Q}$ be the result from running Alg. 1 with inputs $P^*(\mathbf{v}) = L_1(\mathcal{M}^*) > 0$, $\mathcal{G}$, and Q. Then Q is identifiable from $\mathcal{G}$ and $P^*(\mathbf{v})$ if and only if $\widehat{Q}$ is not `FAIL`. Moreover, if $\widehat{Q}$ is not `FAIL`, then $\widehat{Q} = P^{\mathcal{M}^*}(\mathbf{y} \mid do(\mathbf{x}))$.* ■

4 The Neural Estimation Problem

While identifiability is fully solved by the asymptotic theory discussed so far (i.e., both necessary and sufficient), we now focus on the problem of practically estimating causal effects under imperfect optimization and finite samples and computation. We consider the discrete case with binary variables for concreteness, but our construction extends naturally to categorical and continuous variables (see Appendix B). We construct a $\mathcal{G}$-constrained NCM $\widehat{M}(\mathcal{G}; \boldsymbol{\theta}) = \langle \widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, P(\widehat{\mathbf{U}}) \rangle$ (following Def. 7), which leads to the following partially specified NCM:

$$\begin{cases} \mathbf{V} & := \mathbf{V},\ \widehat{\mathbf{U}} := \{U_{\mathbf{C}} : \mathbf{C} \in C^2(\mathcal{G})\} \cup \{G_{V_i} : V_i \in \mathbf{V}\}, \\ \widehat{\mathcal{F}} & := \left\{ f_{V_i} := \arg\max_{j \in \{0,1\}} g_{j,V_i} + \begin{cases} \log \sigma(\phi_{V_i}(\mathbf{pa}_{V_i}, \mathbf{u}^c_{V_i}; \theta_{V_i})) & j = 1 \\ \log(1 - \sigma(\phi_{V_i}(\mathbf{pa}_{V_i}, \mathbf{u}^c_{V_i}; \theta_{V_i}))) & j = 0 \end{cases} \right\}, \\ P(\widehat{\mathbf{U}}) & := \{U_{\mathbf{C}} \sim \text{Unif}(0,1) : U_{\mathbf{C}} \in \mathbf{U}\} \cup \\ & \quad \{G_{j,V_i} \sim \text{Gumbel}(0,1) : V_i \in \mathbf{V}, j \in \{0,1\}\}, \end{cases} \tag{2}$$

where $\mathbf{V}$ are the nodes of $\mathcal{G}$; $\sigma : \mathbb{R} \to (0,1)$ is the sigmoid activation function; $C^2(\mathcal{G})$ is the set of C^2-components of $\mathcal{G}$; each G_{j,V_i} is a standard Gumbel random variable [19]; each $\phi_{V_i}(\cdot; \theta_{V_i})$ is a neural net parameterized by $\theta_{V_i} \in \boldsymbol{\theta}$; $\mathbf{pa}_{V_i}$ are the values of the parents of V_i; and $\mathbf{u}^c_{V_i}$ are the values of $\mathbf{U}^c_{V_i} := \{U_{\mathbf{C}} : U_{\mathbf{C}} \in \mathbf{U} \text{ s.t. } V_i \in \mathbf{C}\}$. The parameters $\boldsymbol{\theta}$ are not yet specified and must be learned through training to enforce L_1-consistency (Def. 4).

Let $\mathbf{U}^c$ and $\mathbf{G}$ denote the latent C^2-component variables and Gumbel random variables, respectively. To estimate $P^{\widehat{M}}(\mathbf{v})$ and $P^{\widehat{M}}(\mathbf{y} \mid do(\mathbf{x}))$ given Eq. 2, we may compute the probability mass of a datapoint $\mathbf{v}$ with intervention $do(\mathbf{X} = \mathbf{x})$ ($\mathbf{X}$ is empty when observational) as:

$$P^{\widehat{M}(\mathcal{G};\boldsymbol{\theta})}(\mathbf{v} \mid do(\mathbf{x})) = \mathop{\mathbb{E}}_{P(\mathbf{u}^c)} \left[\prod_{V_i \in \mathbf{V} \setminus \mathbf{X}} \tilde{\sigma}_{v_i} \right] \approx \frac{1}{m} \sum_{j=1}^{m} \prod_{V_i \in \mathbf{V} \setminus \mathbf{X}} \tilde{\sigma}_{v_i}, \tag{3}$$

where $\tilde{\sigma}_{v_i} := \begin{cases} \sigma(\phi_i(\mathbf{pa}_{V_i}, \mathbf{u}^c_{V_i}; \theta_{V_i})) & v_i = 1 \\ 1 - \sigma(\phi_i(\mathbf{pa}_{V_i}, \mathbf{u}^c_{V_i}; \theta_{V_i})) & v_i = 0 \end{cases}$ and $\{\mathbf{u}^c_j\}_{j=1}^m$ are samples from $P(\mathbf{U}^c)$. Here, we assume $\mathbf{v}$ is consistent with $\mathbf{x}$ (the values of $X \in \mathbf{X}$ in $\mathbf{v}$ match the corresponding ones of $\mathbf{x}$).

Otherwise, $P^{\widehat{M}(\mathcal{G};\boldsymbol{\theta})}(\mathbf{v} \mid do(\mathbf{x})) = 0$. For numerical stability of each $\phi_i(\cdot)$, we work in log-space and use the log-sum-exp trick.

Alg. 1 (lines 2-3) requires non-trivial evaluations of expressions like $\arg\max_{\boldsymbol{\theta}} P^{\widehat{M}}(\mathbf{y} \mid do(\mathbf{x}))$ while enforcing L_1-consistency. Whenever only finite samples are available $\{\mathbf{v}_k\}_{k=1}^n \sim P^*(\mathbf{V})$, the parameters of an L_1-consistent NCM may be estimated by minimizing data negative log-likelihood:

$$\boldsymbol{\theta} \in \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{P^*(\mathbf{v})}\left[-\log P^{\widehat{M}(\mathcal{G};\boldsymbol{\theta})}(\mathbf{v})\right]$$
$$\approx \arg\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{k=1}^{n} -\log \widehat{P}_m^{\widehat{M}(\mathcal{G};\boldsymbol{\theta})}(\mathbf{v}_k). \quad (4)$$

To simultaneously maximize $P^{\widehat{M}}(\mathbf{y} \mid do(\mathbf{x}))$, we subtract a regularization term $\log \widehat{P}_m^{\widehat{M}}(\mathbf{y} \mid do(\mathbf{x}))$, resulting in the objective $\mathcal{L}(\{\mathbf{v}_k\}_{k=1}^n)$ equal to

$$\frac{1}{n}\sum_{k=1}^{n} -\log \widehat{P}_m^{\widehat{M}}(\mathbf{v}_k) - \lambda \log \widehat{P}_m^{\widehat{M}}(\mathbf{y} \mid do(\mathbf{x})), \quad (5)$$

where λ is initially set to a high value and decreases during training. For minimizing, we instead subtract $\lambda \log(1 - \widehat{P}_m^{\widehat{M}}(\mathbf{y} \mid do(\mathbf{x})))$.

Alg. 2 is one possible way of optimizing the parameters $\boldsymbol{\theta}$ required in lines 2,3 of Alg. 1. Eq. 5 is amenable to optimization through standard gradient descent tools, e.g., [32, 43, 42]. [10] [11]

Algorithm 2: Training Model

Input : Data $\{\mathbf{v}_k\}_{k=1}^n$, variables $\mathbf{V}, \mathbf{X} \subseteq \mathbf{V}$, $\mathbf{x} \in \mathcal{D}_{\mathbf{X}}$, $\mathbf{Y} \subseteq \mathbf{V}$, $\mathbf{y} \in \mathcal{D}_{\mathbf{Y}}$, causal diagram $\mathcal{G}$, number of Monte Carlo samples m, regularization constant λ, learning rate η

1 $\widehat{M} \leftarrow \texttt{NCM}(\mathbf{V}, \mathcal{G})$ `// from Def. 7`
2 Initialize parameters $\boldsymbol{\theta}_{\min}$ and $\boldsymbol{\theta}_{\max}$
3 **for** $k \leftarrow 1$ **to** n **do**
  `// Estimate from Eq. 3`
4 $\hat{p}_{\min} \leftarrow \texttt{Estimate}(\widehat{M}(\boldsymbol{\theta}_{\min}), \mathbf{V}, \mathbf{v}_k, \emptyset, \emptyset, m)$
5 $\hat{p}_{\max} \leftarrow \texttt{Estimate}(\widehat{M}(\boldsymbol{\theta}_{\max}), \mathbf{V}, \mathbf{v}_k, \emptyset, \emptyset, m)$
6 $\hat{q}_{\min} \leftarrow 0$
7 $\hat{q}_{\max} \leftarrow 0$
8 **for** $\mathbf{v} \in \mathcal{D}_{\mathbf{V}}$ **do**
9 **if** $\texttt{Consistent}(\mathbf{v}, \mathbf{y})$ **then**
10 $\hat{q}_{\min} \leftarrow \hat{q}_{\min} +$ $\texttt{Estimate}(\widehat{M}(\boldsymbol{\theta}_{\min}), \mathbf{V}, \mathbf{v}, \mathbf{X}, \mathbf{x}, m)$
11 $\hat{q}_{\max} \leftarrow \hat{q}_{\max} +$ $\texttt{Estimate}(\widehat{M}(\boldsymbol{\theta}_{\max}), \mathbf{V}, \mathbf{v}, \mathbf{X}, \mathbf{x}, m)$
  `// ℒ from Eq. 5`
12 $\mathcal{L}_{\min} \leftarrow -\log \hat{p}_{\min} - \lambda \log(1 - \hat{q}_{\min})$
13 $\mathcal{L}_{\max} \leftarrow -\log \hat{p}_{\max} - \lambda \log \hat{q}_{\max}$
14 $\boldsymbol{\theta}_{\min} \leftarrow \boldsymbol{\theta}_{\min} + \eta \nabla \mathcal{L}_{\min}$
15 $\boldsymbol{\theta}_{\max} \leftarrow \boldsymbol{\theta}_{\max} + \eta \nabla \mathcal{L}_{\max}$

One way of understanding Alg. 1 is as a search within the $\Omega(\mathcal{G})$ space for two NCMs that minimizes/maximizes the interventional distribution, $\boldsymbol{\theta}^*_{\min}$ and $\boldsymbol{\theta}^*_{\max}$. Whenever the optimization ends, we can compare the corresponding $P(\mathbf{y} \mid do(\mathbf{x}))$ and determine whether an effect is identifiable. With perfect optimization and unbounded resources, identifiability entails the equality between these two quantities. In practice, we rely on a hypothesis testing step such as

$$|f(\widehat{M}(\boldsymbol{\theta}_{\max})) - f(\widehat{M}(\boldsymbol{\theta}_{\min}))| < \tau \quad (6)$$

for quantity of interest f and a certain threshold τ. This threshold is somewhat similar to the significance level found in statistics, and can be used to control certain types of errors. In our case, the threshold τ can be determined empirically. For further discussion, see Appendix B.

5 Experiments

We first evaluate NCMs regarding their ability to decide whether an effect is identifiable through Alg. 2. Observational data is generated from 8 different SCMs and the corresponding causal diagrams are shown Fig. 4 (top part). Appendix B provides further details of the parametrizations. The datasets along with the causal diagrams are given as inputs to the algorithm to decide whether an effect is identifiable. The NCM does not have access to the true SCM. The target causal effect is $P(Y \mid do(X))$, and the quantity we optimize is the *average treatment effect* (ATE) of X on Y, $ATE_{\mathcal{M}}(X, Y) = \mathbb{E}_{\mathcal{M}}[Y \mid do(X = 1)] - \mathbb{E}_{\mathcal{M}}[Y \mid do(X = 0)]$. Note that if the outcome Y is binary, as in our examples, $\mathbb{E}[Y \mid do(X = x)] = P(Y = 1 | do(X = x))$. The effect is identifiable through do-calculus in the settings represented by Fig. 4 in the left part, and not identifiable in right.

[10] Our approach is flexible and may take advantage of these different methods depending on the context, and there are a number of alternatives for minimizing the discrepancy between P^* and $P^{\widehat{M}}$, including minimizing divergences, such as maximum mean discrepancy [18] or kernelized Stein discrepancy [41], performing variational inference [8], or generative adversarial optimization [15].

[11] The NCM can be extended to the continuous case by replacing the Gumbel-max trick on $\sigma(\phi_i(\cdot))$ with a model that directly computes a probability density given a data point, e.g., normalizing flow [52] or VAE [33].

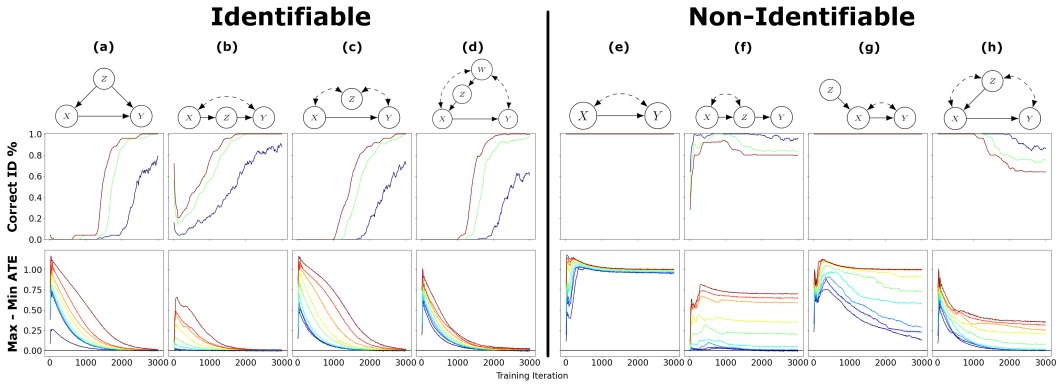

Figure 4: Experimental results on detecting identifiability with NCM. **Top**: Graphs from left to right: (ID cases) back-door, front-door, M, napkin; (not ID cases) bow, extended bow, IV, bad M. **Middle**: Classification accuracy over 3000 training epochs from running hypothesis test on Eq. 6 with $\tau = 0.01$ (blue), 0.03 (green), 0.05 (red). **Bottom**: (1, 5, 10, 25, 50, 75, 90, 95, 99)-percentiles for max-min gaps over 3000 training epochs.

We record the *max-min gaps*, the l.h.s of Eq. 6 with $f(\mathcal{M}) = \text{ATE}_{\mathcal{M}}(X, Y)$, over 3000 training epochs. The parameter λ is set to 1 at the beginning, and decreases logarithmically over each epoch until it reaches 0.001 at the end of training.

Using the hypothesis testing procedure described in Appendix B, these gaps can be used to classify the quantity as "ID" or "non-ID". The gaps and classification accuracy across the 3000 training epochs are shown in Fig. 4. Note that in identifiable settings, the gaps slowly reduce to 0, while the gaps rapidly grow and stay high throughout training in the unidentifiable ones. The classification accuracy for ID cases then gradually increases as training progresses, while accuracy for non-ID cases remain high the entire time (perfect in the bow and IV cases).

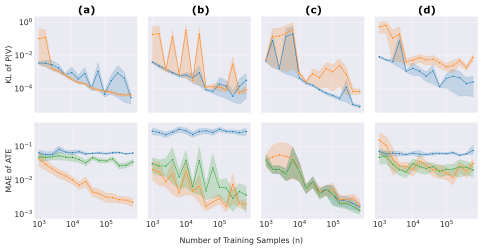

Figure 5: NCM estimation results for ID cases. Columns a, b, c, d correspond to the same graphs as a, b, c, d in Fig. 4. **Top**: KL divergence of $P(\mathbf{V})$ induced by naïve model (blue) and NCM (orange) compared to $P^{\mathcal{M}^*}(\mathbf{V})$. **Bottom**: MAE of ATE of naïve model (blue), NCM (orange), and WERM (green). Plots in log-log scale.

In the identifiable settings, we evaluate the performance of the NCM at estimating the correct causal effect, shown in Fig. 5. As a generative model, the NCM is capable of generating samples from both $P(\mathbf{V})$ and identifiable L_2 distributions like $P(Y \mid do(X))$. We compare the NCM to a naïve generative model trained via likelihood maximization fitted on $P(\mathbf{V})$ without using $\mathcal{G}$ as an inductive bias. Since the naïve model is not defined to sample from $P(Y \mid do(X))$, this shows the implications of arbitrarily choosing $P(Y \mid do(X)) = P(Y \mid X)$. Both models improve at fitting $P(\mathbf{V})$ with more samples, but the naïve model fails to learn the correct ATE except in case (c), where $P(Y \mid do(X)) = P(Y \mid X)$. Further, the NCM is competitive with WERM [29], a state-of-the-art estimation method that directly targets estimating the causal effect without generating samples.

6 Conclusions

We introduced neural causal models (NCMs), which are a special type of SCMs trainable through deep learning techniques. We showed that despite being as expressive as the SCM (Thm. 1), the NCM cannot perform cross-layer inference in general (Corol. 1). Disentangling expressivity and learnability, we formally introduced a new inductive bias based on structural constraints (as encoded in causal diagrams) that empowers NCMs to solve problems in causal inference, including identification and estimation (Thm. 4). We illustrated this by providing a sound and complete training procedure (Alg. 1, 2, Cor. 4) to show how to accomplish this in practice; experimental results support our claim.

References

[1] John E. Angus. The probability integral transform and related results. *SIAM Review*, 36(4):652–654, 1994.

[2] Lawrence J. Appel, Thomas J. Moore, Eva Obarzanek, William M. Vollmer, Laura P. Svetkey, Frank M. Sacks, George A. Bray, Thomas M. Vogt, Jeffrey A. Cutler, Marlene M. Windhauser, and et al. A clinical trial of the effects of dietary patterns on blood pressure. *New England Journal of Medicine*, 336(16):1117–1124, Apr 1997.

[3] Alexander Balke and Judea Pearl. Counterfactual Probabilities: Computational Methods, Bounds, and Applications. In R. Lopez de Mantaras and D.˜Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 46–54. Morgan Kaufmann, San Mateo, CA, 1994.

[4] E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. In R. M. Shiffrin, editor, *Proceedings of the National Academy of Sciences*, volume 113, pages 7345–7352. National Academy of Sciences, 2016.

[5] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On Pearl's Hierarchy and the Foundations of Causal Inference. Technical Report R-60, Causal AI Lab, Columbia University, Also, In "Probabilistic and Causal Inference: The Works of Judea Pearl" (ACM Turing Series), in press, 2020.

[6] Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2015.

[7] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[8] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, Apr 2017.

[9] Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21865–21877. Curran Associates, Inc., 2020.

[10] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, December 1989.

[11] Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *Data Mining and Knowledge Discovery*, May 2021.

[12] Andrew Forney, Judea Pearl, and Elias Bareinboim. Counterfactual Data-Fusion for Online Reinforcement Learners. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[13] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 881–889, Lille, France, 07–09 Jul 2015. PMLR.

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.

[16] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, David Lopez-Paz, Isabelle Guyon, and Michèle Sebag. Learning Functional Causal Models with Generative Neural Networks. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Springer Series on Challenges in Machine Learning. Springer International Publishing, 2018.

[17] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1764–1772, Bejing, China, 22–24 Jun 2014. PMLR.

[18] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19, pages 513–520. MIT Press, 2007.

[19] E.J. Gumbel. *Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures*. Applied mathematics series. U.S. Government Printing Office, 1954.

[20] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. A survey of learning causality with data. *ACM Computing Surveys*, 53(4):1–37, Sep 2020.

[21] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257, 1991.

[22] A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9551–9561, Vancouver, Canada, Jun 2020. Curran Associates, Inc.

[23] A. Jaber, J. Zhang, and E. Bareinboim. Causal identification under Markov equivalence. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pages 978–987. AUAI Press, Aug 2018.

[24] A. Jaber, J. Zhang, and E. Bareinboim. Causal identification under Markov equivalence: Completeness results. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2981–2989, Long Beach, CA, 2019. PMLR.

[25] Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 3020–3029. JMLR.org, 2016.

[26] Y. Jung, J. Tian, and E. Bareinboim. Estimating causal effects using weighting-based estimators. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020. AAAI Press.

[27] Y. Jung, J. Tian, and E. Bareinboim. Learning causal effects via weighted empirical risk minimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12697–12709, Vancouver, Canada, Jun 2020. Curran Associates, Inc.

[28] Y. Jung, J. Tian, and E. Bareinboim. Estimating identifiable causal effects through double machine learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, number R-69, Vancouver, Canada, Feb 2021. AAAI Press.

[29] Yonghan Jung, Jin Tian, and Elias Bareinboim. Learning Causal Effects via Weighted Empirical Risk Minimization. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, 2020.

[30] Nathan Kallus. DeepMatch: Balancing deep covariate representations for causal inference using adversarial training. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5067–5077. PMLR, 13–18 Jul 2020.

[31] Andrej Karpathy. pytorch-made. `https://github.com/karpathy/pytorch-made`, 2018.

[32] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[33] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[34] M. Kocaoglu, A. Jaber, K. Shanmugam, and E. Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14346–14356, Vancouver, Canada, 2019. Curran Associates, Inc.

[35] M. Kocaoglu, K. Shanmugam, and E. Bareinboim. Experimental design for learning causal graphs with latent variables. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 7018–7028. Curran Associates, Inc., 2017.

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc., 2012.

[37] S. Lee and E. Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8565–8576, Vancouver, Canada, Jun 2020. Curran Associates, Inc.

[38] Sanghack Lee and Elias Bareinboim. Structural Causal Bandits: Where to Intervene? In *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems*, 2018.

[39] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861 – 867, 1993.

[40] Sheng Li and Yun Fu. Matching on balanced nonlinear representations for treatment effects estimation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 929–939. Curran Associates, Inc., 2017.

[41] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 276–284, New York, New York, USA, 20–22 Jun 2016. PMLR.

[42] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[44] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6449–6459, Red Hook, NY, USA, 2017. Curran Associates Inc.

[45] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6231–6239. Curran Associates, Inc., 2017.

[46] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*. 2013.

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

[48] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2000.

[49] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018.

[50] Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. Complete Graphical Characterization and Construction of Adjustment Sets in Markov Equivalence Classes of Ancestral Graphs. *Journal of Machine Learning Research*, 18, 2018.

[51] Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[52] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

[53] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[54] Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2503–2513, 2019.

[55] P Spirtes, C N Glymour, and R Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

[56] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[57] Jin Tian. Identifying Conditional Causal Effects. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 561–568. AUAI Press, 2004.

[58] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 2633–2643. Curran Associates, Inc., 2018.

[59] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

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

[61] Junzhe Zhang and Elias Bareinboim. Non-Parametric Methods for Partial Identification of Causal Effects. Technical Report Technical Report R-72, Columbia University, Department of Computer Science, New York, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The bottom of Sec. 1 lists all of our claims along with the corresponding theorem, corollary, or algorithm that supports that claim.
   (b) Did you describe the limitations of your work? [N/A] The limitations of this work are largely derived from the assumptions, but these assumptions (e.g. availability of causal diagram) are universal in causal effect identifiability.
   (c) Did you discuss any potential negative societal impacts of your work? [No] As a well established multidisciplinary field, causal inference has no particular well-known negative social impacts. It is possible that someone may naïvely label some non-causal pattern as causal based on some training results, but this is precisely the kind of mistake to which our paper hopes to shed light.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] For the theoretical results, we mainly assume that $\mathcal{G}$ is provided, and learning is perfect. Finer details are explained in Appendix A.
   (b) Did you include complete proofs of all theoretical results? [Yes] In Appendix A due to space constraints.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Link and instructions are in Appendix B.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Every experiment was run several times with seeds generated deterministically for reproducibility. Details are in Appendix B.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] The only existing asset we used was Andrej Karpathy's implementation of MADE, which we appropriately cited in Appendix B.
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] All of our assets are contained on Github linked in Appendix B.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]