On the Tractability of Neural Causal Inference

Matej Zečević
TU Darmstadt

Devendra Singh Dhami
TU Darmstadt

Kristian Kersting
TU Darmstadt, hessian.AI

Abstract

Roth (1996) proved that any form of marginal inference with probabilistic graphical models (e.g., Bayesian Networks) will at least be NP-hard. Introduced and extensively investigated in the past decade, the neural probabilistic circuits known as sum-product network (SPN) offers linear time complexity. On another note, research around neural causal models (NCM) recently gained traction, demanding a tighter integration of causality for machine learning. To this end, we present a theoretical investigation of if, when, how and under what cost tractability occurs for different NCM. We prove that SPN-based causal inference is generally tractable, opposed to standard MLP-based NCM. We further introduce a new tractable NCM-class that is efficient in inference and fully expressive in terms of Pearl’s Causal Hierarchy. Our comparative empirical illustration on simulations and standard benchmarks validates our theoretical proofs.

1 Introduction

Causal interactions stand at the center of human cognition thus being of high value to science, engineering, business, and law (Penn and Povinelli 2007). Questions like ”What if?” and ”Why?” were discovered to central to how children explore as recent strides in developmental psychology suggest (Gopnik 2012; Buchsbaum et al. 2012; Pearl and MacKenzie 2018), similar to the scientific method. Whereas artificial intelligence research dreams of an automatation to the scientist’s manner (McCarthy 1998; McCarthy and Hayes 1981; Steinruecken et al. 2019), deep learning’s advance brought universality in approximation i.e., for any function there will exist a neural network that is close in approximation to arbitrary precision (Cybenko 1989; Hornik 1991). The field has seen tremendous progress ever since, see for instance (Krizhevsky et al. 2012; Mnih et al. 2013; Vaswani et al. 2017). Thereby, the integration of causality with deep learning is crucial for achieving human-level intelligence. Preliminary attempts, for so-called neural-causal models (Xia et al. 2021; Pawlowski et al. 2020; Zečević et al. 2021a) exist and show to be promising.

While causality has been thoroughly formalized within the last decade (Pearl 2009; Peters et al. 2017), and deep learning advanced, the issue of tractability of inference (Cooper 1990; Roth 1996; Choi et al. 2020) has been left unscathed. It is generally known that semantic graphs like Bayesian Networks (BNs) (Pearl 1995) scale exponentially for marginal inference, while computation graphs (or probabilistic circuits) like sum-product networks (SPNs) (Poon and Domingos 2011) scale in polynomial time. A conversion method considered in Zhao et al. 2015 showed how to compile back and forth between SPNs and BNs. Yet, diverging views on tractable causal inference were reported, see (Papantonis and Belle 2020) and (Zečević et al. 2021a). The former argues using the aforementioned conversion scheme, which leads to a degenerate BN with no causal semantics, while the latter proposes a partial neural-causal model that leverages existing interventional data to tractably perform causal inferences. Motivated by these discrepancies and lack of clarity, this work focuses on investigating systematically if, when, how and under what cost the different types of causal inference occur in tractable manner.

We make the following contributions: (1) We provide theoretical results on SPN-based causal inference, (2) we prove that tractability of inference is being preserved for causal queries, (3) we provide a new SPN-based NCM called TNCM (4) we prove and empirically show that NCM (Xia et al. 2021) are inefficient in comparison TNCM, and (5) we perform an empirical evaluation density and causal effect estimation.

Code available at: https://github.com/zecevic-matej/Tractable-Neural-Causal-Model
2 Background and Related Work

Let us briefly review the background on both the main tractable model class of concern, sum-product networks, as well as key concepts from causality.

Sum-Product Networks. Introduced by [Poon and Domingos (2011)], which generalized the notion of network polynomials based on indicator variables \( \lambda_{X=x} \) for (finite-state) RVs \( X \) from [Darwiche (2003)], sum-product networks (SPN) represent a special type of probabilistic model that allows for a variety of exact and efficient inference routines. SPNs are considered as directed acyclic graphs (DAG) consisting of product, sum and leaf (or distribution) nodes whose structure and parameterization can be efficiently learned from data to allow for efficient modelling of joint probability distributions \( p(X) \). Formally a SPN \( \mathcal{S} = (G, w) \) consists of non-negative parameters \( w \) and a DAG \( G = (V, E) \) with indicator variable \( \lambda \) leaf nodes and exclusively internal sum and product nodes given by,

\[
S(\lambda) = \sum_{C \in \text{ch}(S)} w_{S,C} C(\lambda) \quad P(\lambda) = \prod_{C \in \text{ch}(S)} C(\lambda),
\]

where the SPN output \( S(\lambda) \) is computed at the root node \( \langle S(\lambda) \rangle = S(x) \) and the probability density for \( x \) is \( p(x) = \frac{S(x)}{\sum_{x' \in X} S(x')} \). They are members of the family of probabilistic circuits [Van den Broeck et al. (2019)].

A special class, to be precise, that satisfies properties known as completeness and decomposability. Let \( N \) denote a node in SPN \( \mathcal{S} \), then

\[
\text{sc}(N) = \begin{cases} \{X\} & \text{if } N \text{ is IV } (\lambda_{X=x}) \\ \bigcup_{C \in \text{ch}(N)} \text{sc}(C) & \text{else} \end{cases}
\]

is called the scope of \( N \) and

\[
\forall S \in \mathcal{S} : (\forall C_1, C_2 \in \text{ch}(S) : \text{sc}(C_1) = \text{sc}(C_2))
\]

\[
\forall P \in \mathcal{S} : (\forall C_1, C_2 \in \text{ch}(S) : \ldots \text{C}_1 \neq C_2 \implies \text{sc}(C_1) \cap \text{sc}(C_2) = \emptyset)
\]

are the completeness and decomposability properties respectively. Since their introduction, SPNs have been heavily studied such as by [Trapp et al. (2019)] that present a way to learn SPNs in a Bayesian realm whereas [Kalra et al. (2018)] learn SPNs in an online setting. Several different types of SPNs have also been studied such as Random SPN [Peñaranz et al. (2020b)], Credal SPNs [Levray and Bele, 2020] and Sum-Product-Quotient Networks [Shariat and Shashua (2018)] to name a few. For more, readers are referred to the survey of [Paris, Sánchez-Cauce, and Diez (2020)]. More recently, on the intersection of machine learning and causality, [Zečević et al. (2021a)] proposed an extension to the conditional (or gated) SPN (CSPN) [Shao et al. (2019)] capable of adhering to interventional queries. Formally, an iSPN is being defined as

\[
\mathcal{I} = (\theta, \Psi, S'_{\Psi} : V \rightarrow [0,1])
\]

being a special case to the CSPN-formulation, that is, consider the general formulation of a CSPN \( \mathcal{C} = (\theta, S) \) modelling a conditional distribution \( p_c(Y | X) \) with feed-forward neural network \( \theta : X \rightarrow \Psi \) and SPN \( S_{\Psi} : Y \rightarrow [0,1] \). By realizing that an intervention \( \text{do}(x) \) comes with the mutilation of the causal graph \( G = (V,E) \) such that new graph is \( G' = (V, \{ (i,j) : (i,j) \in E \land i \notin \text{pa}(X) \}) \), the iSPN is able to formulate an intervention for SPN natural to the occurrence of interventions in structural causal model. The gate model \( g \) orchestrates the \( do \)-queries such that the density estimator (SPN) can easily switch between different interventional distributions.

Causal Inference. A Structural Causal Model (SCM) as defined by [Peters et al. (2017)] is specified as \( \mathcal{C} := (\mathcal{S}, P(U)) \) where \( P(U) \) is a product distribution over exogenous unmodelled variables and \( \mathcal{S} \) is defined to be a set of \( d \) structural equations

\[
V_i := f_i(\text{pa}(V_i), U_i), \quad \text{where } i = 1, \ldots, d
\]

with \( \text{pa}(V_i) \) representing the parents of variable \( V_i \) in graph \( G(\mathcal{C}) \). An intervention \( \text{do}(W), W \subset V \) on a SCM \( \mathcal{C} \) as defined in \( \mathcal{I} \) occurs when (multiple) structural equations are being replaced through new non-parametric functions \( g_W \) thus effectively creating an alternate SCM \( \mathcal{C}_2 := \mathcal{C}^{\text{do}(W=g_W)} \). Interventions are referred to as "imperfect" if the parental relation is kept intact, \( g_i(\text{pa}(i), \cdot) \), and as atomic if \( g_i = a \) for \( a \in \mathbb{R} \).

An important property of interventions often referred to as "modularity" or "autonomy"\(^1\) states that interventions are fundamentally of local nature, formally

\[
p_{\mathcal{C}_2}(V_i | \text{pa}(V_i)) = p_{\mathcal{C}_1}(V_i | \text{pa}(V_i)),
\]

where the intervention of \( \mathcal{C}_2 \) occurred on variable \( V_j \) opposed to \( V_i \). This suggests that mechanisms remain invariant to changes in other mechanisms which implies that only information about the effective changes induced by the intervention need to be compensated for. An important consequence of autonomy is the truncated factorization

\[
p(V) = \prod_{V \notin W} p(V | \text{pa}(V))
\]

derived by [Pearl (2009)], which suggests that an intervention \( \text{do}(W) \) introduces an independence of a set of intervened nodes \( W \) to its causal parents. Another

\(^1\)See Section 6.6 in [Peters et al. (2017)].
important assumption in causality is that causal mechanisms do not change through intervention suggesting a notion of invariance to the cause-effect relations of variables which further implies an invariance to the origin of the mechanism i.e., whether it occurs naturally or through means of intervention (Pearl et al., 2016). A SCM $C$ is capable of emitting various mathematical objects such as graph structure, statistical and causal quantities placing it at the heart of causal inference, rendering it applicable to machine learning applications in marketing (Hair Jr and Sarstedt, 2021), healthcare (Bica et al., 2020) and education (Holles and Schaar, 2016). A SCM induces a causal graph $G$, an observational/associational distribution $p$, an interventional distribution $p_{\text{do}(\ldots)}$, can be intervened upon using the $\text{do}$-operator and thus generate interventional distributions $p_{\text{do}(\ldots)}$ and given some observations $v$ can also be queried for interventions within a system with fixed noise terms amounting to counterfactual distributions $p_{v=\text{do}(\ldots)}$. As suggested by the Causal Hierarchy Theorem (CHT) (Bareinboim et al., 2020), these properties of an SCM almost always form the Pearl Causal Hierarchy (PCH) consisting of different levels of distributions being $L_1$ associational, $L_2$ interventional and $L_3$ counterfactual. This hierarchy suggests that causal quantities $(L_i, i \in \{2, 3\})$ are in fact richer in information than statistical quantities $(L_1)$, and the necessity of causal information (e.g. structural knowledge) for inference based on lower rungs e.g. $L_1 \not\rightarrow L_2$. Finally, to query for samples of a given SCM, the structural equations are being simulated sequentially following the causal hierarchy of endogenous variables $V$. To conclude, consider the formal definition of valuations for the first two layers being

$$p^g(v,|do(x)) = \sum_{\{u:Y_u=x\}} p(u) \quad \text{(9)}$$

for node sets and instances $X, Y, x, y$ where $Y_u:U \rightarrow Y$ denotes the value of $Y$ under intervention $x$.

3 Causal Inference with SPN

Our theoretical analysis starts off with the relation between SPN and Causality. More specifically, we consider how causal inference occurs within the tractable machinery of SPNs. Subsequently, the definition of the iSPN as a tractable neural model for causal estimation follows naturally.

3.1 Estimation and Vanilla SPN

To expand further on the boundaries of the integration between causality and machine learning, we perform a theoretical inspection on how causal inference can occur with (in) SPN. Such an investigation is important since assuming the wrong causal structure or ignoring it altogether could be fatal w.r.t. any form of generalization out of data support as suggested in (Peters et al., 2017). Central to said (assumed) causality is the concept of intervention. Although being a wrong statement as suggested by results on identifiability, the famous motto of Peter Holland and Don Rubin ‘No causation without manipulation’ (Holland, 1986) phrases interventions as the core concept in causality. In agreement with this view that distributional changes present in the data due to experimental circumstances need be accounted for, we focus our analysis on queries $Q = p(y|\text{do}(x))$ on the second (interventional) level $L_2$ of the PCH (Pearl and Mackenzie, 2018) (Bareinboim et al., 2020), acts as our guiding reference that illustrates the idea behind interventional estimation on a real-world inspired data set for the iSPN model (Eq.5) to be developed in the subsequent sections.

We first define the concept of a statistical estimand $(L_1)$ for SPN as the application of the rules of probability theory (and Bayes Theorem) to the induced joint distribution.

Definition 1. (SPN modelling.) Let $S$ be a SPN with joint distribution $p_S(x) \in L_1$. Then any aggregation in terms of sums $\sum$ or products $\prod$ of conditionals $p(v|w)$ and marginals $p(v)$ for $V, W \subseteq X$ derivable from $p_S$ is being referred to as SPN-estimand.

Note that for a general SPN-estimand $q$ (Def.1) to actually be estimable from data, full probabilistic support $(\forall x: p(x) > 0)$ needs to be assumed since otherwise the estimate might be undefined. Following, we provide our first important theoretical insight in that
Proposition 1. (Interventional SPN-estimand.)
Let \( Q \in \mathcal{L}_2 \) be an identifiable query. There exists an SPN-estimand \( q \) such that \( Q = q \).

Due to space constraints, we provide only a few identified results, thereby transferring tractability of inference also to causal inference.

Corollary 1. (SPN-identification.)
Let \( G \) be the causal graph of a Markovian SCM \( \mathcal{C} \) for which we query \( Q \in \mathcal{L}_2 \) with SPN-estimand \( q \) i.e., \( Q = q \). Then there exists an estimand based on the do-calculus \( \text{do}^{-}\)-calculus \( \text{do}^{-}\)-calculus (Pearl, 2009), denoted as \( g \), such that \( Q = g = q \).

While unsurprising from a causal viewpoint, from the perspective of tractable models research the results in Prop.\( [\text{I}] \) and Cor.\( [\text{I}] \) provide a new incentive for research on the integration of both fields. Papantonis and Belle (2020) considered the usage of the SPN-BN compilation method from (Zhao et al., 2015) for causal inference within SPN that failed due to the resulting BN being a bipartite graph in which the variables of interest were not connected (connectivity being crucial to non-trivial causal inference). To reap initial rewards, we now prove that causal inference with SPN is tractable.

Theorem 1. (Tractable Causal Inference.)
Let \( q \) be the interventional estimand of SPN \( \mathcal{S} = ((V, E), w) \) for \( Q \in \mathcal{L}_2 \). Further, let \( p = |q| \) and \( r = |E| \), with \( p < r \), denote the sizes of the estimand and network respectively. Then \( q \) is being computed in time linear in the size of the network \( O(r) \).

Opposed to (causal) BN where inference is generally intractable (\( \#P \) complexity), Thm.\( [\text{I}] \) suggests that any estimand can be computed efficiently using SPN even if the estimand identifies an interventional quantity, thereby transferring tractability of inference also to causal inference.

3.2 SPN-based Causal Models
An important restriction of SPN-based causal inference is that the joint distribution \( p_{\theta}(v) \) of SPN \( \mathcal{S} \) optimizes all possibly derivable distributions, thereby diminishing single distribution expressivity. I.e., any causal inference will hold but actual estimation from data will suffer in quality. In addition, the positive support assumption might render some practical inference undefined. Therefore, in the following we extend our theoretical analysis to the extension of the SPN model class itself. More specifically, we consider interventional SPN (iSPN) firstly introduced by (Zeccevič et al., 2021a). Our first observation is that the iSPN allows for a compressed model description over the SCM, while trading in expressivity since the iSPN has no means of computing the highest level of the PCH, \( \mathcal{L}_3 \) (counterfactuals). The iSPN (Eq.\( [\text{III}] \)) is more powerful than the SPN by construction, therefore we observe the following.

Proposition 2. (iSPN-estimand Transformation.)
Let \( \mathcal{J} \) and \( \mathcal{S} \) be an iSPN and SPN respectively. There exists a graph \( G \) for which any SPN-estimand \( q_{\mathcal{S}} \) can be modelled with an iSPN-estimand \( q_{\mathcal{J}} \).

Prop.\( [\text{II}] \) suggests that iSPN are indeed joint density estimators, although being defined as a special case of conditional estimators (CSPN), and that any SPN will be covered by the observational distribution \( (\mathcal{L}_1) \) of a corresponding iSPN. In the following, assuming corresponding data \( D_i \sim p_{\mathcal{L}_2} \), we prove that iSPN allow for direct causal estimation of the interventional query \( (\mathcal{L}_2) \) opposed to the previous base SPN causal inference from Prop.\( [\text{I}] \) where the indirect statistical estimand \( (\mathcal{L}_1) \) is acquired through identification (e.g., do-calculus (Pearl, 2009)).

Proposition 3. (Interventional iSPN-estimand)
Let \( Q = p(y|\text{do}(x)) \in \mathcal{L}_2 \) be an identifiable query. Let \( G' \) denote the mutilated graph upon intervention \( \text{do}(x) \) on the SCM \( \mathcal{C} \) with causal graph \( G \). Then from Prop.\( [\text{I}] \) in (Zeccevič et al., 2021a) it follows that there exists an iSPN-estimand \( q_{G'} \) such that \( q_{G'} = Q \).

To illustrate the difference between the parallel approaches of causal inference based on SPN, that is SPN- (Prop.\( [\text{I}] \) and iSPN-based (Prop.\( [\text{III}] \), consider the following example,

\[
\mathcal{C} := \{f_X(Z, U_X), f_Y(X, Z, U_Y), f_Z(U_Z), p(U)\}
\]

where the identification equality is given by the backdoor-adjustment formula on \( \mathcal{C} \) (Pearl, 2009; Peters et al., 2017). The l.h.s. will be modelled by iSPN, while the r.h.s. consisting of multiple terms will be modelled by the SPN. Thereby, the shortcomings of single distribution expressivity and positive support are being resolved. Upon establishing direct causal learning using iSPN, we now answer the question on tractability of causal inference.

Theorem 2. (TCI with iSPN.)
Let \( q_{G'} \) be the interventional estimand of iSPN \( \mathcal{J} = (q_{\theta}, \mathcal{S}_{\theta}) \) for \( Q = p(y|\text{do}(x)) \in \mathcal{L}_2 \) using the mutilated graph \( G' \).
3.3 Tractable Neural Causal Model

In the previous subsections we discussed vanilla SPN-based causal inference, that is tractable (Thm.2), but that requires an identification engine to acquire the actual estimand (e.g. do-calculus), and we discussed iSPN-based causal inference, that too is tractable (Thm.3), but that requires access to intervention data (essentially trading it for the identification engine). Now, in the following, we will move onto more general causal models and theoretically investigate tractability of causal inference for these more complex models. For this, consider a recent stride in neural-causal based methods ignited by the theoretical findings in (Xia et al., 2021), where the authors introduced a parameterized SCM with neural net function approximators to model each of the SCM’s structural equations. Instead of considering neural nets, we will consider the more general instance of an SCM $C(\Theta, \Phi) = \{S(\Theta), P(U)\}$ parameterized by arbitrary choice of function approximator $f_\theta, \theta \in \Theta, i \in \{1,\ldots, |S|\}$ but assuming the same type of approximator, e.g., neural net, is being applied to all the structural equations in $C$. Note that the formulation of a SCM with its structural equations implies a graph, like a Bayesian Network (BN), and this graph is not a computation graph like a SPN but a semantic one in which edges denote causal relations. Unfortunately, it turns out, this heritage of a (parameterized) SCM leads to its intractability for causal (marginal) inference. We state the following result.

Theorem 3. (Intractability of SCM.) Let $C(\Theta) = \{S(\Theta), P(U)\}$ be a parameterized SCM where each structural equation $f_\theta, \theta \in \Theta, i \in \{1,\ldots, |S|\}$ and $D_i$ denotes a variable’s set. Marginal causal inference scales exponentially in the number of variables $O(\prod_i D_i) = O(2^{|S|})$.

From a computational perspective, the result in Thm.3 is a protest against the original formulation of the SCM in terms of practicality. Although being an arguably simple consequence of the BN-heritage of the SCM, still, Thm.3 strongly advises against any efforts of using parameterized SCM for real-world impact. Even if the parameterization comes from powerful approximators like neural nets - causal inference remains intractable. However, for both the sake of completion and the interest of establishing the theoretical connection in the scope of this systematic investigation, we present for the first time a new parameterization of the NCM using SPN. This is sensible since any partial inference within the parameterized SCM might still be efficient - thus SPN can still offer a more pragmatic alternative while not compromising on performance since the functions that are being modelled by the structural equations might often times be simple mechanisms due to their local and restricted nature - thereby neural nets cannot leverage their universal application.

Historically, the SCM developed from the Causal BN which developed from the BN (Pearl, 1995, 2009).
approximation capabilities while SPN can still leverage tractability. Therefore, we now present the Tractable Neural Causal Model (TNCM) formally.

**Definition 2. (Tractable NCM.)** Like in Thm.1 let $\mathfrak{S}$ be a parameterized SCM. If each structural equation $f_\theta \in \mathfrak{S}$ is represented through an SPN $\mathcal{S}_i = ((V, E), \theta_i)$, $\mathcal{S}_i = f_\theta$, then we call $\mathfrak{S}$ a Tractable NCM.

In Fig.2 we provide a schematic comparison of the two causal models based on SPN units i.e., iSPN (Zečević et al., 2021a) and TNCM (Def.2). Evidently, the TNCM is concerned with a more complex model description, yet because of that, it becomes a causal model fully expressive in terms of the PCH. We now state the simple consequence of defining an SCM with SPN units instead of neural nets.

**Corollary 2. (Efficiency Comparison.)** Let $\mathfrak{M} = (\mathfrak{S}_V, p(\mathfrak{U}))$ be an NCM (Xia et al., 2021) where each $f_\theta \in \mathfrak{S}_V$ is a feed-forward neural network and let $\mathfrak{S}$ be a TNCM (Def.2) where each $f_i$ is an SPN of size $r$. Any inference based on $Q \in \mathcal{L}_2$ will be at least quadratic ($n \geq 2$) in the hidden layer size $O(h^n)$ for the NCM and linear in the network size $O(r)$ for the TNCM.

Cor.2 suggests that restricted causal inference (e.g. not marginal inference) even with NCM is tractable but inefficient when compared to TNCM since the former has quadratic (in the case of vector-valued $V_i \in \mathbb{R}^n$ even cubic, and depending on the estimand possibly quartic) time complexity opposed to linear for the latter. Said comparison behaves the same for the iSPN, since Thm.2 suggests that (for a fixed iSPN state) any inference will also be linear. Further extending the comparison to other neural-causal models as suggested by (Zečević et al., 2021b), we find that for the NCM-

### Algorithm 1 Causal Inference with TNCM

**Input:** SCM $\mathfrak{C}$, Obs. $\mathbf{v} \sim \mathcal{E}_1(\mathfrak{C})$, Interv. $do(\mathbf{x})$

**Parameter:** Monte Carlo samples $m$

**Output:** $p(\mathbf{v}| \text{do}(\mathbf{x}))$

1: Let $\mathfrak{T} \leftarrow \text{TNCM}((\mathbf{V}, G(\mathfrak{C})), \Theta)$ (Def.2)
2: if $\mathbf{X} = \emptyset$ then
3:   $p \leftarrow \min(\mathfrak{T}(\mathbf{v}, m, \emptyset), 1)$
4: else if $\mathbf{X} \neq \emptyset \land \text{Consistency}(\mathbf{x}, \mathbf{v})$ then
5:   $p \leftarrow \min(\mathfrak{T}(\mathbf{v}, m, \mathbf{x}), 1)$
6: else
7:   return $0$
8: $p \leftarrow p \ast 1_{V_i=0} + (1-p) \ast 1_{V_i=1}$
9: end if
10: return $p$

Type 2 time complexity is worse being cubic since modelling occurs on edge- opposed to structural equation level. For the iVGAE, which is comparable to the iSPN in terms of model description, the time complexity is as bad as for the NCM. Conclusively, iSPN (Eq.5) offer a clear advantage over other neural-causal models in terms of inference efficiency since any causal query will be answered in linear time, whereas NCM-variants and causal BNs have worse time complexities (quadratic/cubic and exponential, respectively). However, it is important to note that NCM-variants might offer for more expressivity in terms of the PCH.

Thereby, a researcher might choose one model over the other based on the specific application of interest (e.g. efficacy versus expressivity). Fig.3 offers a schematic illustration of how an interventional quantity is being computed for a TNCM (Def.2) and iSPN (Eq.5) in contrast to an SCM (or NCM), clearly highlighting the different in terms of both computation and model description. Upon investigating various causal applications of SPN, we offer a conclusive overview in Tab.1 on the tractability of their neural-causal inferences.

### 3.4 Training and Estimation with TNCM

Since TNCM are a special case of NCM with SPN as parameterizing units, we can apply inference in the same way. I.e., we make use of the truncated factorization formula (Pearl, 2009), more specifically, a sample-based approximation thereof,

$$p(\mathbf{v}| \text{do}(\mathbf{x})) \approx \frac{1}{m} \sum_{i=1}^{m} \prod_{x \in \mathbf{V} \setminus \mathbf{X}} f_{\theta_i}(\mathbf{v}),$$

where $m$ is the number of samples for the unmodelled/noise terms $U_i$. The intuition behind this formula is that an intervention will mutilate the original causal graph deleting dependence on $\mathbf{x}$’s parents. To perform training, one can simply resort to the
Suggest capabilities in modelling the final PCH layer, the tracability of the computation of any single sub-module (i.e., structural equation). Finally, Counterfactuals estimation quality on interventional distribution mod-
elling? is the number of data points. We provide an algorithm pseudo-code sketch (see Alg.1). The consistency criterion refers to the assumption that a query like \( p(y = 1, x = 1 \mid \text{do}(x = 0)) \) should automatically evaluate to zero.

### 4 Empirical Illustration

To give empirical support of the theoretical results on tractability for the various SPN-based approaches to causal inference, we conduct mainly two types of experiments. First, we assess their quality in performing sensible causal inference i.e., not resorting to simple conditional estimates in cases where the equality of conditions and interventions does not hold. Second, we assess the actual inference time complexity for the discussed methods.

More specifically, we answer the following questions: **Q1:** To which degree are causal effects being captured on qualitatively different structures? **Q2:** How is the estimation quality on interventional distribution modelling? **Q3:** How does time complexity scale when increasing the SCM size? **Q4:** How do different methods compare when the the SCM mechanism representation is being scaled?

**Data Sets.** Since we are interested in qualitative behavior in light of the theoretical results established previously, we consider custom SCM simulations. For instance consider the following two models: the collider SCM \( \mathcal{C}_1 \) given by

\[
\begin{align*}
X & \leftarrow f_X(W, U_X) = W \land U_X \\
Y & \leftarrow f_Y(U_Y) = U_Y \\
Z & \leftarrow f_Z(X, Y, U_Z) = X \lor (Y \land U_Z) \\
W & \leftarrow f_W(U_W) = U_W
\end{align*}
\]

and the backdoor SCM \( \mathcal{C}_4 \) given by

\[
\begin{align*}
X & \leftarrow f_X(Z, U_X) = Z \lor U_X \\
Y & \leftarrow f_Y(W, X, U_Y) = X \land (W \land U_Y) \\
Z & \leftarrow f_Z(U_Z) = U_Z \\
W & \leftarrow f_W(Z, U_W) = Z \lor U_W,
\end{align*}
\]

where \( \lor, \land, \lor \) denote logical XOR, OR, and AND. The other SCMs (chain and confounder) that are being deployed are being detailed in the Appendix. Note that (for simplicity of analysis) we consider binary variables, however, (T)NCM naturally extend to the categorical and continuous variables. Note that the collider is an unconfounded structure, thereby conditioning amounts to intervening, \( p(y|x) = p(y|\text{do}(x)) \), while for the backdoor this equality does not hold - thus the causal effect from \( X \) on \( Y \) is confounded via the backdoor \( X \leftarrow \ldots \) over nodes \( Z, W \). We choose \( U \sim \text{Unif}(a, b) \) to be uniform random variables each, and we randomize parameters \( a, b \).

**Protocol and Parameters.** To account for reproducibility and stability of the presented results, we used learned models for four different random seeds each parameterization of any given underlying SCM. For the NCM’s neural networks, we deploy simple MLP with three hidden layers of 10 neurons each, and the input-/output-layers are \(|pa_i| + 1 \) and 1 respectively. For the TNCM’s SPNs, we deploy simple two-layer SPNs (following the layerwise principle introduced in (Peharz et al., 2020a)) where the first layer consists of leaf nodes, the second layer of product nodes, the third layer of sum nodes and a final product node aggregation. The number of channels is set to 30. We use ADAM (Kingma and Ba, 2014) optimization, and train up to three passes of 10k data points sampled from the observational distribution of any SCM. For experiments in which the size of the SCM is being increased, we use a simple chain and extend it iteratively. For experiments in which the capacity of the mechanism (or units) of the parameterized SCM are

| SPN | Only Observations | Implicit-Identification | Tractable Marginals | Tractable Equations | Counterfactuals |
|-----|------------------|------------------------|---------------------|---------------------|----------------|
| iSPN | ✓ | ✗ | ✓ (Thm.1) | - | ✗ |
| NCM | ✓ | ✗ | ✓ (Thm.1) | ✓ (Cor.2) | ✓ |
| TNCM | ✓ | ✓ | ✓ (Thm.1) | ✓ (Cor.2) | ✓ |

Table 1: **Overview on Neural-Causal Inference.** Only Observations refers to the setting where only data from \( L_1 \) is accessible. Implicit-Identification suggest that no external identification engine (e.g. do-calculus) is necessary since identification occurs within the model. Tractable Marginals refer to whether the general computation scheme \( p(x) = \sum_{v \setminus x} p(x, v) \) is computable in polynomial time (tractable). Tractable Equations refers to the tractability of the computation of any single sub-module (i.e., structural equation). Finally, Counterfactuals suggest capabilities in modelling the final PCH layer, \( L_3 \). Pointers are given for explicit results in this work.
On the Tractability of Neural Causal Inference

Figure 4: ATE. Averaged over multiple random seeds over multiple parameterizations of the given SCM. Both NCM and TNCM perform well in estimating causal effects. (Best viewed in color.)

Table 2: Density Estimation. Averaged JSD values. being increased, we use a fixed chain SCM structure and scale the model capacity linearly. I.e., the MLPs increase their hidden layers neurons number while SPNs increase their layer channel. For general causal inference we resort to Alg. 1. For causal effect estimation, we focus on the average treatment effect given by \( ATE(T, E) := E[E|do(T = 1)] - E[E|do(T = 1)] \) that for the binary setting reduces to probabilistic difference \( p(Y = 1|do(X = 1)) - p(Y = 1|do(X = 0)) = ATE(T, E) \). For measuring density estimation quality, we resort to the Jensen-Shannon-Divergence (JSD) with base 2 that is bounded in \([0, 1]\) where 0 indicates identical probability mass functions.

(Q2 Density Estimation, Tab. 2) We observe adequate modelling of the different densities (for the actual plots consider the appendix) since error rates lie mostly in the low single digit domain. Most notably is the increased variance of the \( do(X = 1) \) distribution for TNCM on \( M_1 \). Observing closely, we see that even the other distributions already show less-optimal performance. Since all experiments are conducted with the same, simple architectures, we argue that this non-optimization is explanatory.

(Q3/4 Time Complexities: Graph and Mechanism Size Increase, Fig. 5) We observe the expected intractability of both NCM and TNCM (Thm. 1), while also observing the discrepancy/relativity of the intractability when considering each separate mechanism (or unit) and thus answer the theoretical suggestions affirmatively.

5 Conclusions

Starting from first principles, we investigated three different approaches to causal inference when using SPN providing tractability results and introducing a new tractable NCM (Def. 2) along the way: purely within SPN (Cor. 1, Thm. 1), with partial NCM like iSPN (Prop. 3, Thm. 2), and (T)NCM (Cor. 2). We provide a general result on parametric-SCM that inherit intractability from their semantic (Thm. 1). We provided an empirical illustration showcasing said tractability result, while discussing density and causal effect estimation qualities of TNCM in comparison to NCM.
Acknowledgements

This work was supported by the ICT-48 Network of AI Research Excellence Center “TAILOR” (EU Horizon 2020, GA No 952215) and by the Federal Ministry of Education and Research (BMBF; project “PlexPlain”, FKZ 01IS19081). It benefited from the Hessian research priority programme LOEWE within the project WhiteBox, the HMWK cluster project “The Third Wave of AI” and the Collaboration Lab “AI in Construction” (AICO).

References

E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. On pearl’s hierarchy and. Technical Report, 2020.

I. Bica, A. Alaa, and M. Van Der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In ICML, 2020.

D. Buchsbaum, S. Bridgers, D. Skolnick Weisberg, and A. Gopnik. The power of possibility: Causal learning, counterfactual reasoning, and pretend play. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1599):2202–2212, 2012.

Y. Choi, A. Vergari, and G. Van den Broeck. Probabilistic circuits: A unifying framework for tractable probabilistic models. Technical report, Technical report, 2020.

G. F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks. Artificial intelligence, 42(2-3):393–405, 1990.

G. Cybenko. Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems, 2(4):303–314, 1989.

A. Darwiche. A differential approach to inference in bayesian networks. Journal of the ACM (JACM), 50(3):280–305, 2003.

A. Gopnik. Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. Science, 337(6102):1623–1627, 2012.

J. F. Hair Jr and M. Sarstedt. Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing. Journal of Marketing Theory and Practice, 2021.

W. Hoiles and M. Schaar. Bounded off-policy evaluation with missing data for course recommendation and curriculum design. In ICML, 2016.

P. W. Holland. Statistics and causal inference. Journal of the American statistical Association, 81(396): 945–960, 1986.

K. Hornik. Approximation capabilities of multilayer feedforward networks. Neural networks, 4(2):251–257, 1991.

A. Kalra, A. Rashwan, W.-S. Hsu, P. Poupart, P. Doshi, and G. Trimponias. Online structure learning for feed-forward and recurrent sum-product networks. NeurIPS, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25:1097–1105, 2012.

A. Levray and V. Belle. Learning credal sum-product networks. AKBC, 2020.

J. McCarthy. What is artificial intelligence? 1998.

J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In Readings in artificial intelligence, pages 431–450. Elsevier, 1981.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013.

I. Papantonis and V. Belle. Interventions and counterfactuals in tractable probabilistic models: Limitations of contemporary transformations. arXiv preprint arXiv:2001.10905, 2020.

I. París, R. Sánchez-Cauce, and F. J. Díez. Sum-product networks: A survey. arXiv preprint arXiv:2004.01167, 2020.

N. Pawlowski, D. C. Castro, and B. Glocker. Deep structural causal models for tractable counterfactual inference. arXiv preprint arXiv:2006.06485, 2020.

J. Pearl. From bayesian networks to causal networks. In Mathematical models for handling partial knowledge in artificial intelligence. 1995.

J. Pearl. Causality. Cambridge university press, 2009.

J. Pearl and D. Mackenzie. The book of why: the new science of cause and effect. Basic books, 2018.

J. Pearl, M. Glymour, and N. P. Jewell. Causal inference in statistics: A primer. John Wiley & Sons, 2016.

R. Peharz, S. Lang, A. Vergari, K. Stelzner, A. Molina, M. Trapp, G. V. d. Broeck, K. Kersting, and Z. Ghahramani. Einsum networks: Fast and scalable learning of tractable probabilistic circuits. ICML, 2020a.
R. Peharz, A. Vergari, K. Stelzner, A. Molina, X. Shao, M. Trapp, K. Kersting, and Z. Ghahramani. Random sum-product networks: A simple and effective approach to probabilistic deep learning. In UAI, 2020b.

D. C. Penn and D. J. Povinelli. Causal cognition in human and nonhuman animals: A comparative, critical review. *Annu. Rev. Psychol.*, 58:97–118, 2007.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference*. The MIT Press, 2017.

H. Poon and P. Domingos. Sum-product networks: A new deep architecture. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 689–690. IEEE, 2011.

D. Roth. On the hardness of approximate reasoning. *Artificial Intelligence*, 82(1-2):273–302, 1996.

X. Shao, A. Molina, A. Vergari, K. Stelzner, R. Peharz, T. Liebig, and K. Kersting. Conditional sum-product networks: Imposing structure on deep probabilistic architectures. *arXiv preprint arXiv:1905.08550*, 2019.

O. Sharir and A. Shashua. Sum-product-quotient networks. In *AISTATS*, 2018.

C. Steinruecken, E. Smith, D. Janz, J. Lloyd, and Z. Ghahramani. The automatic statistician. In *Automated Machine Learning*, pages 161–173. Springer, Cham, 2019.

M. Trapp, R. Peharz, H. Ge, F. Pernkopf, and Z. Ghahramani. Bayesian learning of sum-product networks. *NeurIPS*, 2019.

G. Van den Broeck, N. Di Mauro, and A. Vergari. Tractable probabilistic models: Representations, algorithms, learning, and applications. *Tutorial at UAI*, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

K. Xia, K.-Z. Lee, Y. Bengio, and E. Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. *arXiv preprint arXiv:2107.00793*, 2021.

M. Zečević, D. S. Dhami, A. Karanam, S. Natarajan, and K. Kersting. Interventional sum-product networks: Causal inference with tractable probabilistic models. In *NeurIPS*, 2021a.

M. Zečević, D. S. Dhami, P. Veličković, and K. Kersting. Relating graph neural networks to structural causal models. *arXiv preprint arXiv:2109.04173*, 2021b.

H. Zhao, M. Melibari, and P. Poupart. On the relationship between sum-product networks and bayesian networks. In *ICML*, 2015.
Appendix - Tractable Neural Causal Models

We make use of this appendix following the main paper to provide the proofs to the main theorems, propositions, and corollaries in addition to further details on the experiments and all generated density plots.

1 Proofs

Due to space constraints for the main text, we provide the mathematical proofs for the theoretical results here.

1.1 Proof for Theorem 1

A key insight. The following is a general statement on parametric-SCM that suggests they inherit intractability from their semantic nature (i.e., Bayesian networks heritage [Pearl 1995, 2009]).

Theorem 1. (Intractability of SCM.) Let \( \mathcal{C}(\Theta) = \{S(\Theta), P(U)\} \) be a parameterized SCM where each structural equation \( f_\theta i : D_{pa(i)} \times U_i \rightarrow D_i, f_\theta i \in S \) is represented through a parameterized function approximator \( \theta_i \in \Theta \) and \( D_i \) denotes a variable’s set. Marginal causal inference scales exponentially in the number of variables \( O(\prod_i D_i) = O(2^{|S|}) \).

Proof. Ad absurdum. Assume there exists a parameterized SCM \( \mathcal{C}(\Theta) \) that can perform tractable inference. Then \( \mathcal{C} \) computes \( p(x) \) in polynomial time, \( O(\text{poly}(c)) \) for some computation cost \( c \). Decomposing the computation of parametric-SCM \( \mathcal{C} \) using its implied graph structure, we observe \( p(v) = \prod_i p(v_i | pa(v_i)) = \sum_{v \setminus x} p(x,v) \) where latter is an exponential term and not a polynomial. Thus we have a contradiction and there exists no parameterized SCM with tractable causal marginal inference.

1.2 Proofs for Proposition 1 and Corollary 1

The following statement suggest that we can always use an identification engine (e.g. do-calculus) jointly with an SPN estimator.

Proposition 1. (Interventional SPN-estimand.) Let \( Q \in \mathcal{L}_2 \) be an identifiable query. There exists an SPN-estimand \( q \) such that \( Q = q \).

Proof. Let \( Q = p(y | do(x)) \) be the identifiable query to be represented using the SPN \( S \) and its joint distribution \( p_S \). Since \( Q \) is identifiable, \( Q \in \mathcal{L}_2 \) can be written in terms of statistical terms i.e., there exists an aggregation in terms of sums and products \( q^* \in \mathcal{L}_1 \) such that \( Q = q^* \) where \( q^* \) is derivable from the observational joint distribution \( p_E \) of SCM \( \mathcal{C} \). Since a SPN can represent any joint distribution, there exists a \( p_S \) s.t. \( p_S = p_E \).

The following statement is a natural consequence of the previous one. Apart from the existence, one particular choice of identification engine is the do-calculus.

Corollary 1. (SPN-identification.) Let \( G \) be the causal graph of a Markovian SCM \( \mathcal{C} \) for which we query \( Q \in \mathcal{L}_2 \) with SPN-estimand \( q \) i.e., \( Q = q \). Then there exists an estimand based on the application of the do-calculus [Pearl 2009] denoted as \( g \) such that \( Q = g = q \).

Proof. The do-calculus [Pearl 2009] is complete i.e., any identifiable query \( Q \in \mathcal{L}_2 \) will in fact be identified by some application of its three rules i.e., there always exists a do-calculus estimand \( g \) based on graph \( G \) for which \( Q = g \). Since an SPN can represent any joint distribution, we can always find an SPN-estimand \( q \) based on the graphical do-calculus such that \( g = q \).
1.3 Proof for Theorem 1

The following statement suggests that SPN-based causal inference is tractable if the estimand complexity is restricted.

**Theorem 2. (Tractable Causal Inference.)** Let \( q \) be the interventional estimand of SPN \( S = (\mathcal{V}, \mathcal{E}, w) \) for \( Q \in \mathcal{L}_2 \). Further, let \( p=|\mathcal{V}| \) and \( r=|\mathcal{E}| \), with \( p<r \), denote the sizes of the estimand and network respectively. Then \( q \) is being computed in time linear in the size of the network \( \mathcal{O}(r) \).

**Proof.** The SPN \( S \) requires per estimand-term in \( q \) a bottom-up computation linear in the size of the network \( r \), \( \mathcal{O}(r) \) [Poon and Domingos, 2011]. Since \( p<r \), the overall time complexity for evaluating all \( p \) components does not change. Therefore, the causal quantity \( Q \in \mathcal{L}_2 \) is computed in \( \mathcal{O}(r) \).

1.4 Proof for Proposition 2

The following statement suggests that iSPN are more powerful than SPN since they constitute a superset.

**Proposition 2. (Interventional iSPN-estimand Transformation.)** Let \( \mathcal{I} \) and \( S \) be an iSPN and SPN respectively. There exists a graph \( G \) for which any SPN-estimand \( q_S \) can be modelled with an iSPN-estimand \( q^G \).

**Proof.** For any SPN-estimand \( q_S \), there exists an SCM \( \mathcal{C} \) with induced graph \( G \) such that the joint distribution by the SPN \( p_S(x) \) coincides with the observational distribution of the SCM i.e., \( L_1(\mathcal{C}) = p_S \). Since iSPN extend on the regular SPN by allowing an arbitrary causal structure \( G' \), one can select the SCM graph \( G' := G \) such that the iSPN joint distribution \( p_r \) coincides with the SCM and thereby with the original SPN. Estimands are derived from the joint distribution.

1.5 Proof for Proposition 3

As for the vanilla SPN approach, an iSPN also acquires an estimand but through implicit means.

**Proposition 3. (Interventional iSPN-estimand.)** Let \( Q = p(y \mid do(x)) \in \mathcal{L}_2 \) be an identifiable query. Let \( G' \) denote the mutilated graph upon intervention \( do(x) \) on the SCM \( \mathcal{C} \) with causal graph \( G \). Then there exists an iSPN-estimand \( q^{G'} \) such that \( q^{G'} = Q \).

**Proof.** Assuming the data used for the iSPN-optimization to originate from an interventional distribution of SCM \( \mathcal{C} \) with graph \( G \), that is \( v \sim p(v \mid do(x)) \in \mathcal{L}_2 \). Upon evaluation of the parameters of the iSPN \( \mathcal{I}=(g_{\theta}, S_{\psi}) \) induced SPN \( S_{\psi}, \psi = g_{\theta}(G^{do(x)}) \), the modelled joint distribution is an interventional distribution \( p_S(v \mid do(x)) \in \mathcal{L}_2 \). Then, the analogue to the proof for Prop.1 applies.

1.6 Proofs for Theorem 2 and Corollary 2

The following statement suggest that causal inference within iSPN is tractable, yet for different reasons than the SPN-based approach.

**Theorem 3. (TCI with iSPN.)** Let \( q^{G'} \) be the interventional estimand of iSPN \( \mathcal{I} = (g_{\theta}, S_{\psi}) \) for \( Q = p(y \mid do(x)) \in \mathcal{L}_2 \) using the mutilated graph \( G' \). Like in Thm.1, let \( p \) and \( r \) with \( p<r \) denote the sizes of the estimand and network respectively. Any inference based on \( Q \) upon evaluating the parameters of \( S_{\psi} \) via \( \psi = g_{\theta}(G') \) is being computed in time linear in the size of the network \( \mathcal{O}(r) \).

**Proof.** Since any iSPN reduces to an SPN upon parameter-evaluation, we can apply the proof to Thm.1 that any inference from the joint distribution will be linear in network size.

The following statement, following our key insight from Thm.1, suggest that classical NCM are worse off than TNCM.

**Corollary 2. (NCM Inefficiency.)** Let \( \mathcal{M} = (S_V, p(U)) \) be an NCM [Xia et al., 2021] where each \( f_i \in S_V \) is a feed-forward neural network and let \( Q \in \mathcal{L}_2 \). Assuming \( f_i \) of same architecture where \( i = |V|, h_1, h_2, o = 1 \) denote input-, hidden-, and output-layer sizes with \( h_1 = h_2 > i \). Any inference based on \( Q \) will be quadratic in the hidden layer size \( \mathcal{O}(h^2) \).
Proof. Any feed-forward neural network $f$ requires a forward pass that involves a matrix multiplication. For the assumed architecture, this boils down to $f(x) = \sigma_2(W_2\sigma_1(W_1x + b_1) + b_2)$ with $\sigma_i, b_i,$ and $W_i$ being non-linearities, biases, and weights respectively. Since $h > i$, but any SCM variable $V_i \in \mathbb{R}$ is real, this operation scales quadratically in the hidden layer size $O(ih^2) = O(h^2)$.

2 Other SCM Structures, Code, and Density Plots

Remaining SCM Structures. We considered two more SCM structures. A simple chain which has no confounding given by

$$c_2 = \left\{ \begin{array}{l} X \leftarrow f_X(U_X) = U_X \\ Y \leftarrow f_Y(X, U_Y) = X \land U_Y \\ Z \leftarrow f_Z(Y, U_Z) = Y \land U_Z \\ W \leftarrow f_W(Z, U_W) = Z \land U_W, \end{array} \right. \tag{11}$$

and finally the confounded structure is given by

$$c_4 = \left\{ \begin{array}{l} X \leftarrow f_X(Z, U_X) = Z \lor U_X \\ Y \leftarrow f_Y(X, Z, U_Y) = (X \land U_Y) \oplus (Z \land U_Y) \\ Z \leftarrow f_Z(U_Z) = U_Z \\ W \leftarrow f_W(X, U_W) = X \land U_W, \end{array} \right. \tag{12}$$

Technical Details and Code. All experiments are being performed on a MacBook Pro (13-inch, 2020, Four Thunderbolt 3 ports) laptop running a 2.3 GHz Quad-Core Intel Core i7 CPU with a 16 GB 3733 MHz LPDDR4X RAM on time scales between a few second and an hour. Our code is available at: [https://github.com/zecevic-matej/Tractable-Neural-Causal-Model](https://github.com/zecevic-matej/Tractable-Neural-Causal-Model)

Density Plots. Following this page are the density estimations, computed for NCM and TNCM, averaged over the random seeds for each parameterizations of each different SCM and plotted against the ground truth.
Figure 6: NCM. Four random seeds per plot, blue is the estimation, black is the ground truth. Variable order: $X, Y, Z, W$. (Best viewed in color.)
Figure 7: TNCM. Four random seeds per plot, blue is the estimation, black is the ground truth. Variable order: $X, Y, Z, W$. (Best viewed in color.)