Causal Inference

Hedong Yan,
Computer Department, HKBU

Xia, Kevin, Kai-Zhan Lee, Yoshua Bengio, and **Elias Bareinboim**. "The Causal-Neural Connection: Expressiveness, Learnability, and Inference." (2021). NeurIPS2021
Zečević, M., Dhami, D.S., Karanam, A., Natarajan, S. and Kersting, K., 2021. Interventional Sum-Product Networks: Causal Inference with Tractable Probabilistic Models. arXiv preprint arXiv:2102.10440. NeurIPS2021

The connection with my work

My work is about causal inference, individual treatment, and medical AI.

And treatment effect estimation for Medical AI requires strong **explainable/intervention model** and **credible number**.

Paper 1: an **end-to-end** (causal effect identification and estimation) **neural method** for arbitrary L_2 query (estimand)

Paper 2: a **tractable** causal neural model based on conditional sum-product network

Persons

Ronald A. Fisher

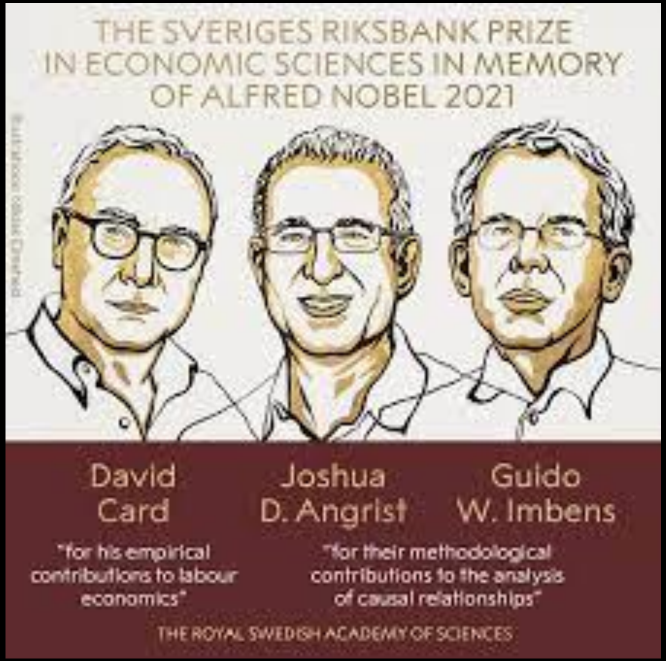
Judea Pearl

Donald Rubin

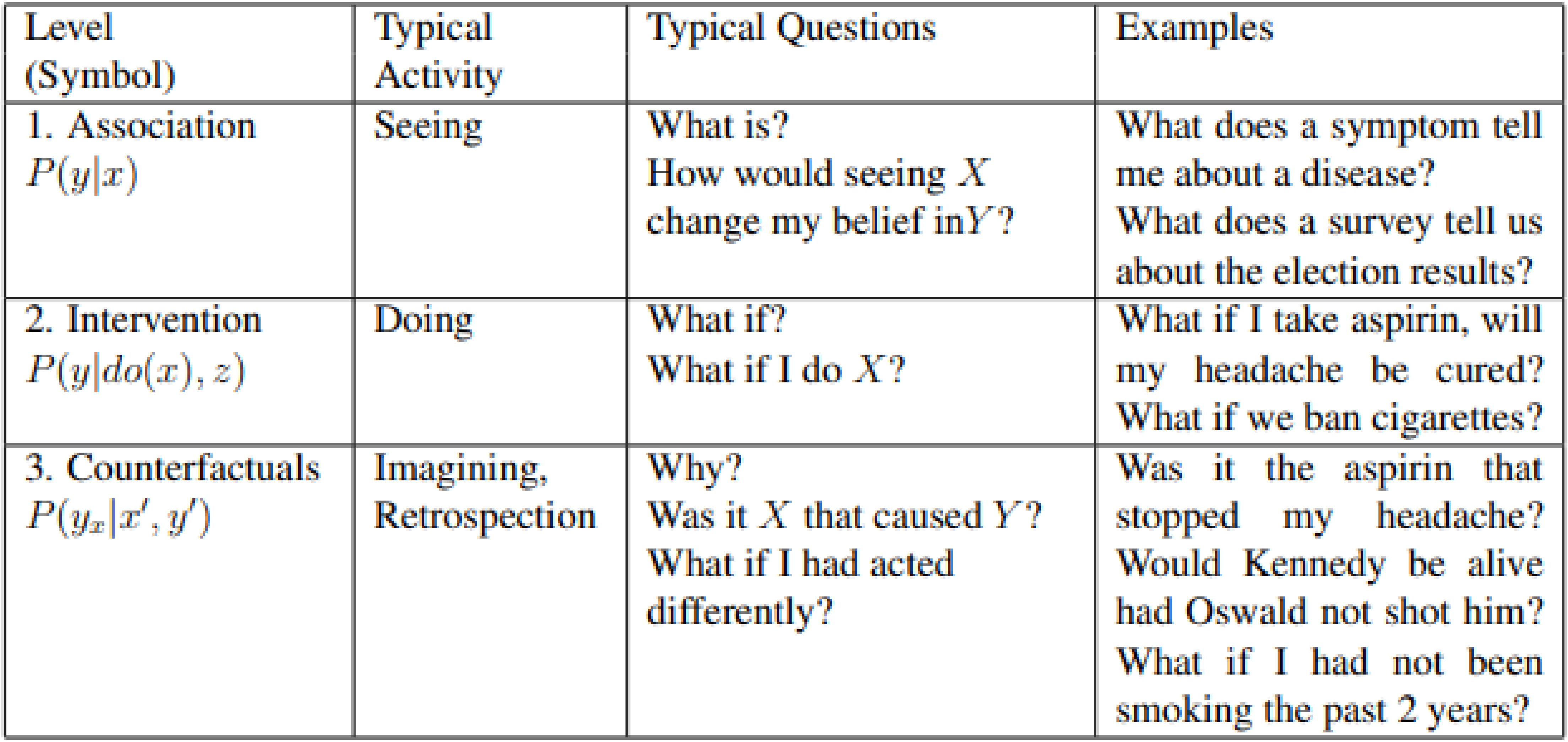
Jerzy Neyman

Elias Bareinboim

Yousua Bengio


THE SVERIGES RIKSBANK PRIZE
IN ECONOMIC SCIENCES IN MEMORY
OF ALFRED NOBEL 2021
David
Card
"for his empirical
contributions to labour
economics"
Joshua
D. Angrist
Guido
W. Imbens
"for their methodological
contributions to the analysis
of causal relationships"
THE ROYAL SWEDISH ACADEMY OF SCIENCES

and more ……



Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in$Y$?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 1: The ladder of causation

Pearl, J., 2000. Models, reasoning and inference. Cambridge, UK: Cambridge University Press, 19.
Pearl, J. and Mackenzie, D., 2018. The book of why: the new science of cause and effect. Basic books.
http://web.cs.ucla.edu/~kaoru/3-layer-causal-hierarchy.pdf

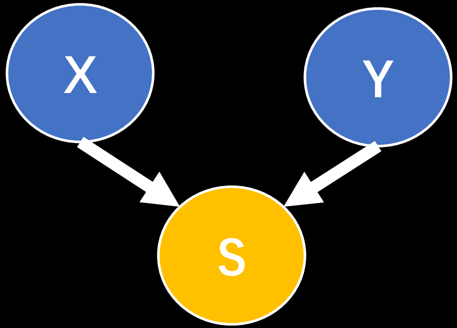
Application

- Recommendation system (A/B, popularity bias etc.)
- CV and NLP (stable learning, IRM, HRM etc.)
- Robotic
- Causal reinforcement learning (POMAP etc.)
- Transfer learning (transportability, data fusion etc.)
- Economics (labor markets, natural experiment etc.)
- Climate

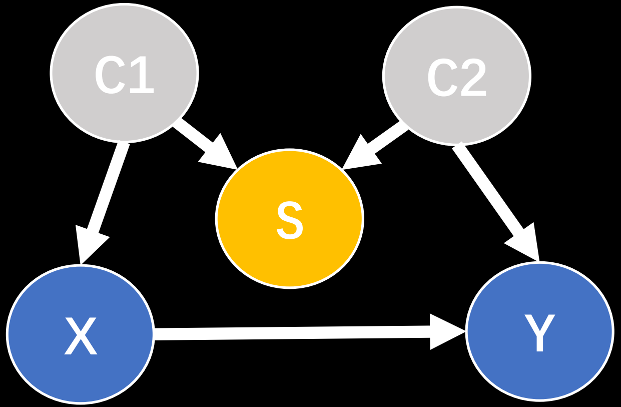
Basic

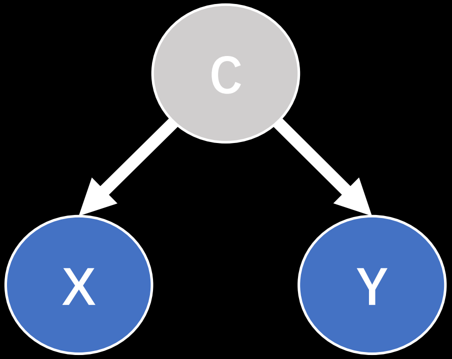
- modularity (modules independence and model/data independence)
- do-operation (mutilation) and do-calculus

Causal effect doesn’t include spurious correction!

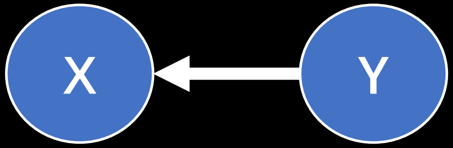


selection bias

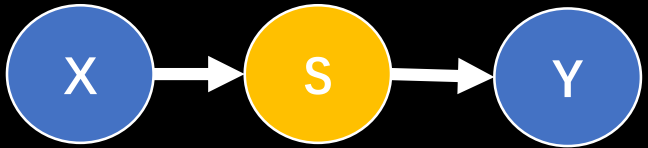




confounding bias



anti-causal



mediation bias

M bias

ZHANG jiji. HKBU. Do-calculus and Modularity in Causal Markov Categories. 2021 PCIC talk.

Basic

Do-operation is an operation to remove arrow/influence to the variable of SCM that we want to do intervention on.

Do-calculus is a COMPLETE and SUFFICIENT algebraic method for ALL identifiable cases to prompt L_2 Query globally with local Markov property of DAGs and give L_0 Response in polynomial time by L_1 $D\mathrm{ata}$.

Pearl, Judea (1995), "Causal diagrams for empirical research", Biometrika, 82 (4): 669–710, doi:10.1093/biomet/82.4.669.
Huang, Yimin; Valtorta, Marco (2006). "Pearl's Calculus of Intervention is Complete". *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*: 217–224.

Paper 1: The Causal-Neural Connection: Expressiveness, Learnability, and Inference

- Motivation: disentangling the notions of expressivity and learnability
- Methods: g-constrained Neural Causal Model
- Experimental dataset: generated by equations
- Major findings: an end-to-end algorithm that is both sufficient and necessary to determine whether a causal effect can be learned from data (causal identifiability) and then estimates the effect whenever identifiability holds (causal estimation/estimand) or give up and low bound

Expressivity and Learnability

Expressivity of Neural Model

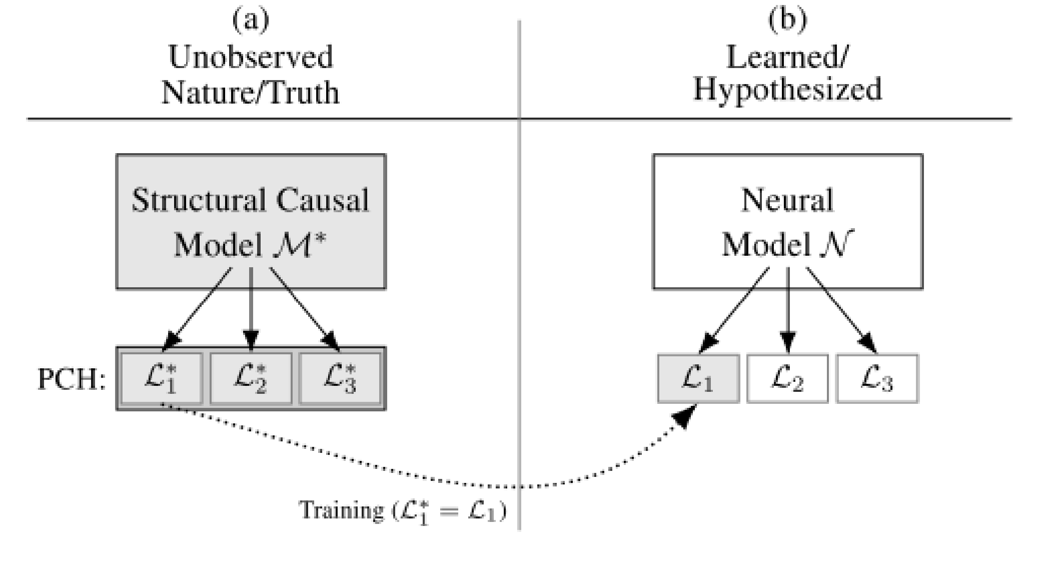


Figure 1: The l.h.s. contains the unobserved true SCM $\mathcal{M}^*$ that induces the three layers of the PCH. The r.h.s. contains an NCM that is trained to match in layer 1. The matching shading indicates that the two models agree w.r.t. L_1 while not necessarily agreeing w.r.t. layers 2 and 3.

Expressivity of Neural Causal Model

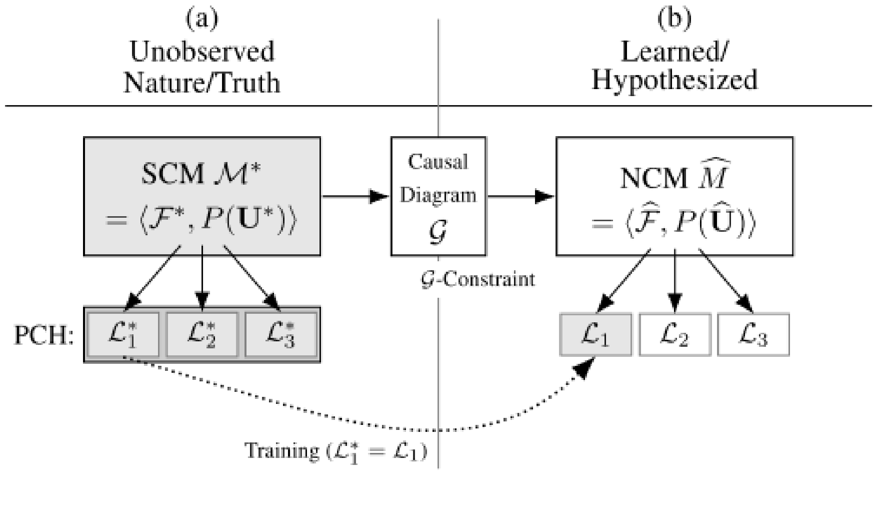
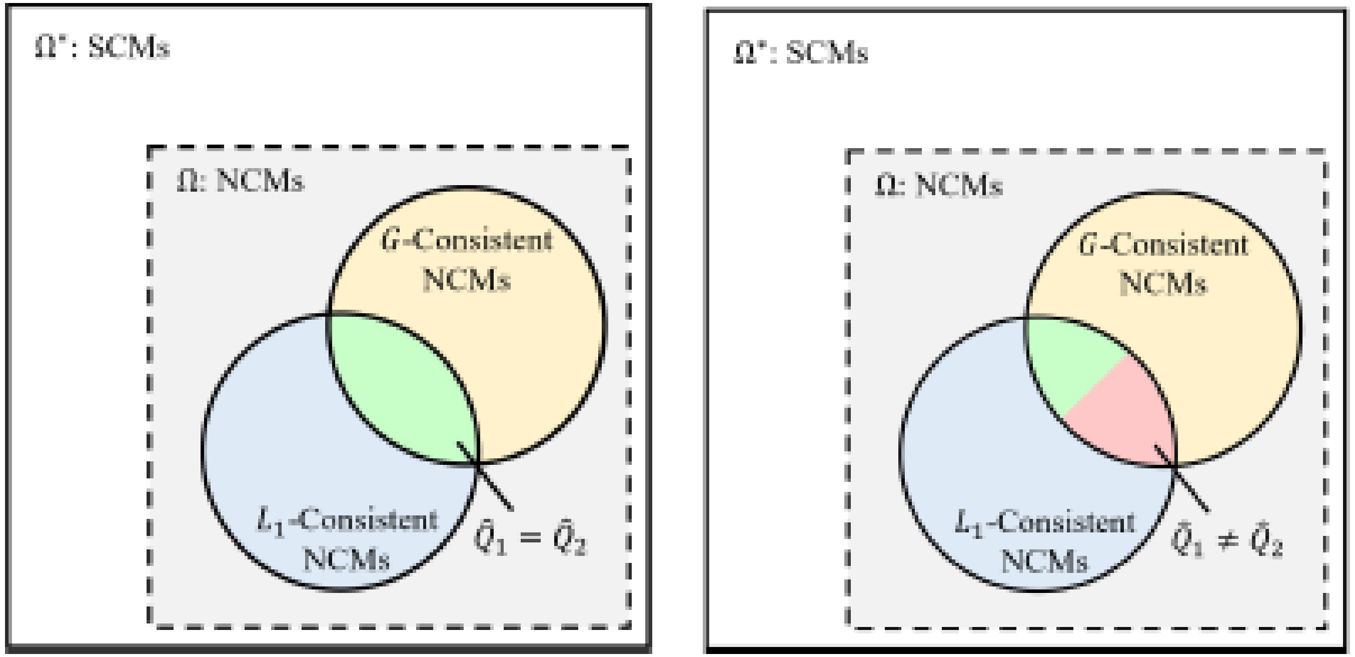


Figure 2: The l.h.s. contains the true SCM $\mathcal{M}^*$ that induces PCH's three layers. The r.h.s. contains an NCM that is trained with layer 1 data. The matching shading indicates that the two models agree with respect to L_1 while not necessarily agreeing in layers 2 and 3. The causal diagram $\mathcal{G}$ entailed by $\mathcal{M}^*$ is used as an inductive bias for $\widehat{M}$.

Expressivity and Learnability

Identifiability



(a) In the identifiable case, all NCMs that are $\mathcal{G}$-consistent and L_1-consistent with $\mathcal{M}^*$ will also match in Q.

(b) In the non-identifiable case, there could exist two NCMs, $\widehat{M}_1$ and $\widehat{M}_2$, that are both $\mathcal{G}$-consistent and L_1-consistent but still disagree in Q.

Figure 11: A visual representation of the ID problem. Here, Q is a query of interest, and $\widehat{Q}_i$ is the answer for that query induced by NCM $\widehat{M}_i$. The goal is to check if all NCMs that are $\mathcal{G}$-consistent and L_1-consistent are also consistent in Q.

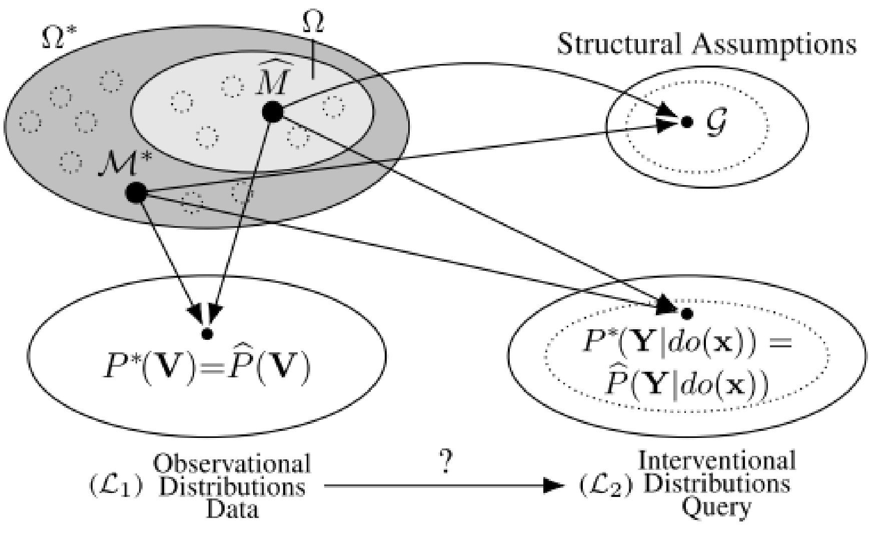
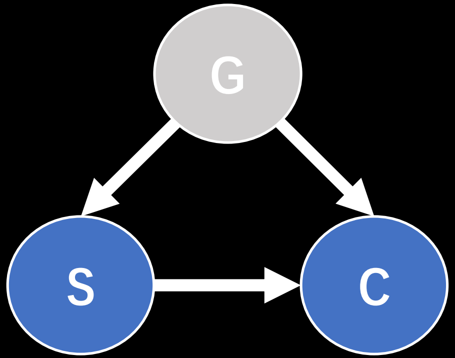
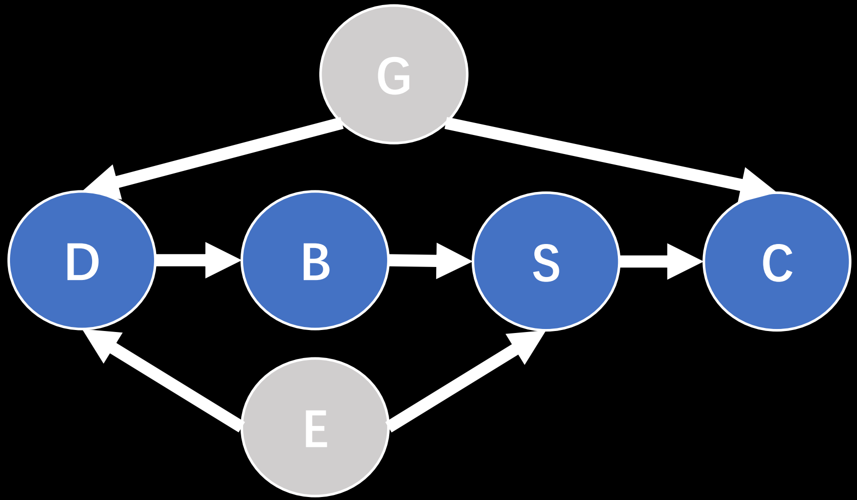

Ω^*
Ω
$\widehat{M}$
$\mathcal{M}^*$
Structural Assumptions
$\mathcal{G}$
$P^*(\mathbf{V})=\widehat{P}(\mathbf{V})$
$P^*(\mathbf{Y}|do(\mathbf{x})) = \widehat{P}(\mathbf{Y}|do(\mathbf{x}))$
$(\mathcal{L}_1)$ Observational Distributions Data
?
$(\mathcal{L}_2)$ Interventional Distributions Query

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).

Example of Identifiability





P(cancer | do(smoke)) is NOT identifiable due to unobserved gene.
M1: G->S and G->C are NOT 0 and S->C is 0
M2: G->S and G->C are 0 AND S->C is NOT 0
It maybe a bad causal graph.

D means dopamine; B means brain; G means undetected gene/physique ; E mean social environment not easy to measure.

P(cancer | do(smoke)) is minimum graph that is identifiable by do-calculus but not identifiable by front-door and back-door.

$$P(c|do(s)) = \frac{\sum_d p(c,s|d,b)p(d)}{\sum_d p(s|d,b)p(d)}$$

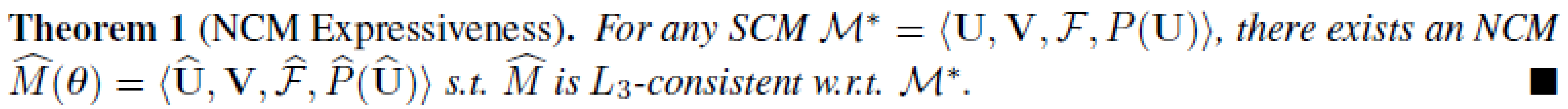
Zečević, M., Dhami, D.S., Karanam, A., Natarajan, S. and Kersting, K., 2021. Interventional Sum-Product Networks: Causal Inference with Tractable Probabilistic Models. arXiv preprint arXiv:2102.10440.

Extra Glance Material

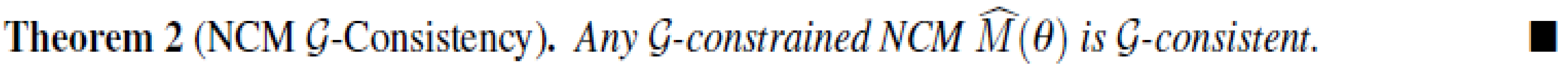
Structural Causal Model:
https://wiki.swarma.org/index.php?title=%E7%BB%93%E6%9E%84%E5%9B%A0%E6%9E%9C%E6%A8%A1%E5%9E%8B

Do-calculus:
https://wiki.swarma.org/index.php?title=Do%E6%BC%94%E7%AE%97

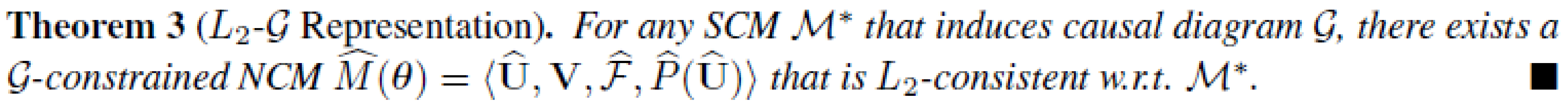
Theorems



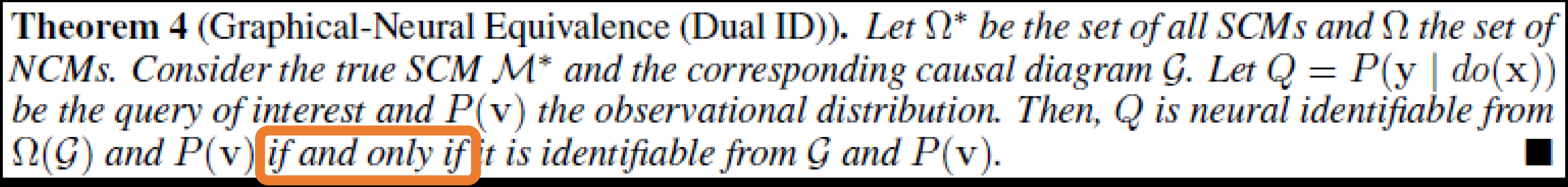
Theorem 1 (NCM Expressiveness). *For any SCM* $\mathcal{M}^* = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, *there exists an NCM* $\widehat{M}(\theta) = \langle \widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, \widehat{P}(\widehat{\mathbf{U}}) \rangle$ *s.t.* $\widehat{M}$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



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

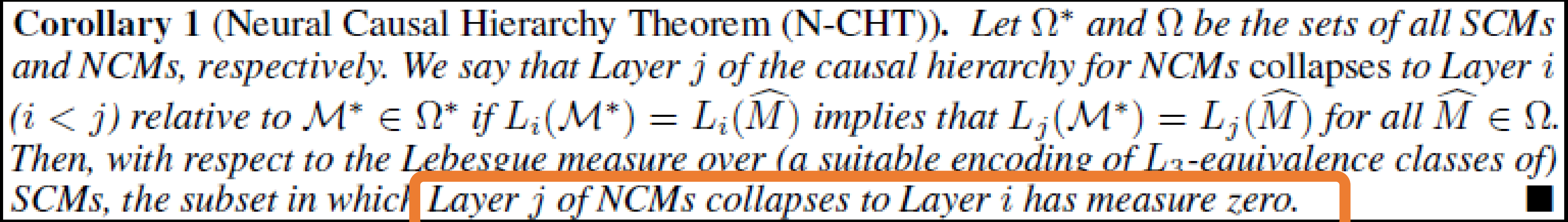


Theorem 3 (L_2-$\mathcal{G}$ Representation). *For any SCM* $\mathcal{M}^*$ *that induces causal diagram* $\mathcal{G}$, *there exists a* $\mathcal{G}$*-constrained NCM* $\widehat{M}(\theta) = \langle \widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, \widehat{P}(\widehat{\mathbf{U}}) \rangle$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

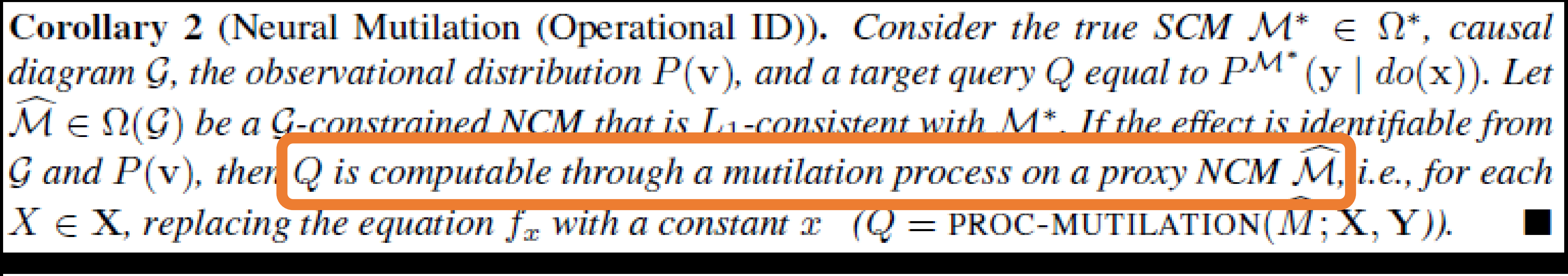


Theorem 4 (Graphical-Neural Equivalence (Dual ID)). *Let* Ω^* *be the set of all SCMs and* Ω *the set of NCMs. Consider the true SCM* $\mathcal{M}^*$ *and the corresponding causal diagram* $\mathcal{G}$. *Let* $Q = P(\mathbf{y} \mid do(\mathbf{x}))$ *be the query of interest and* $P(\mathbf{v})$ *the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

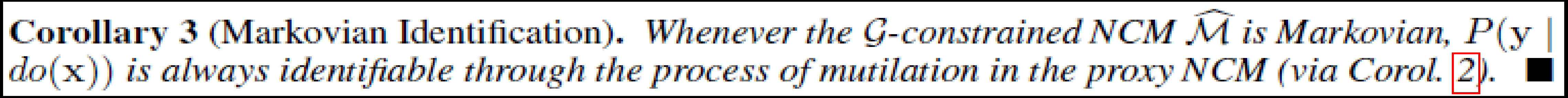
Expressivity and Learnability



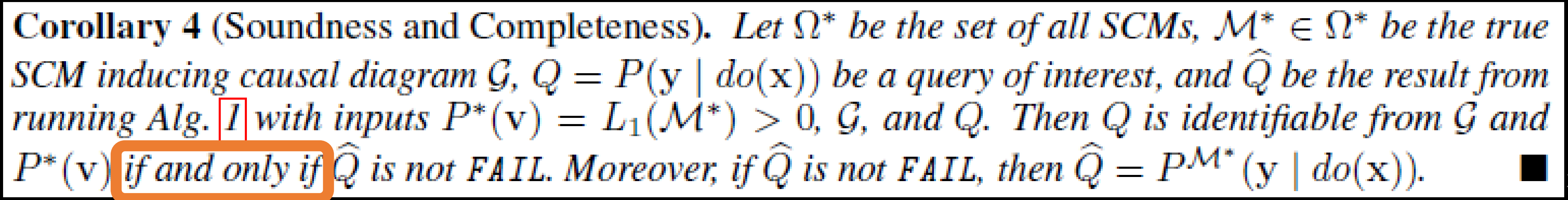
Corollary 1 (Neural Causal Hierarchy Theorem (N-CHT)). *Let Ω^* and Ω be the sets of all SCMs and NCMs, respectively. We say that Layer j of the causal hierarchy for NCMs* collapses *to Layer i $(i < j)$ relative to $\mathcal{M}^* \in \Omega^*$ if $L_i(\mathcal{M}^*) = L_i(\widehat{M})$ implies that $L_j(\mathcal{M}^*) = L_j(\widehat{M})$ 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.* ■



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 $\widehat{M} \in \Omega(\mathcal{G})$ 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 $\widehat{M}$, i.e., for each $X \in \mathbf{X}$, replacing the equation f_x with a constant x $(Q =$ PROC-MUTILATION$(\widehat{M}; \mathbf{X}, \mathbf{Y}))$.* ■

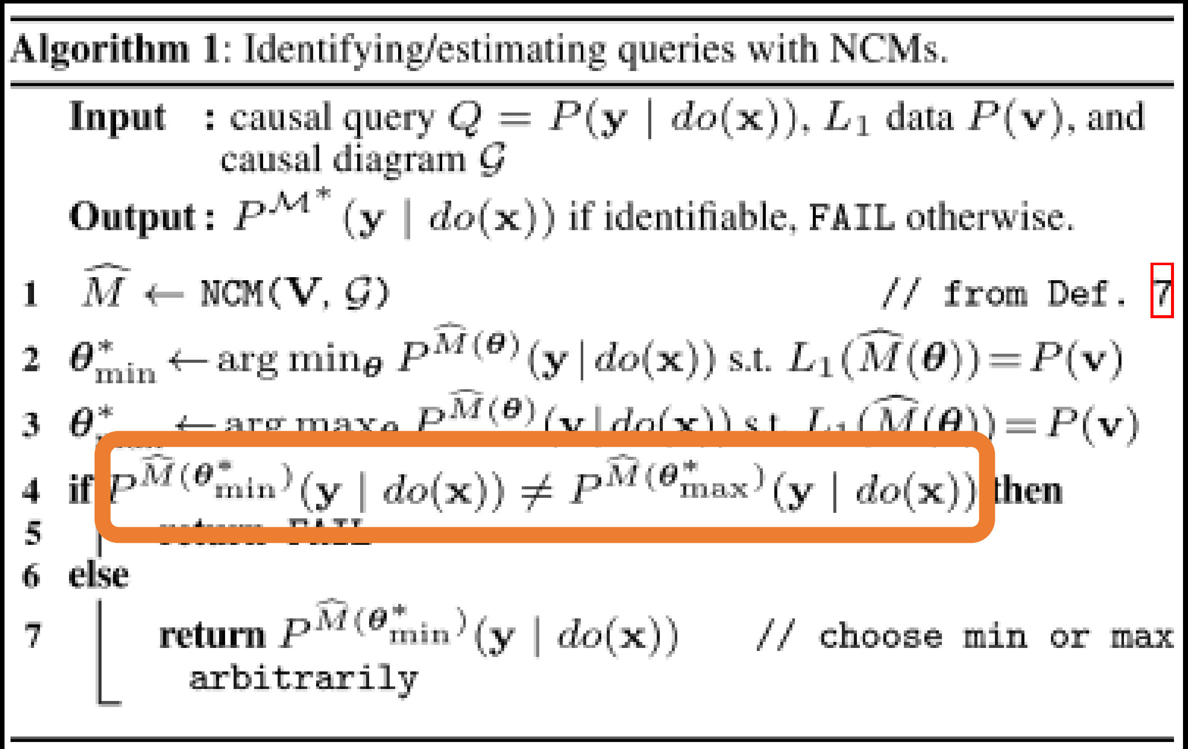


Corollary 3 (Markovian Identification). *Whenever the $\mathcal{G}$-constrained NCM $\widehat{M}$ is Markovian, $P(\mathbf{y} \mid do(\mathbf{x}))$ is always identifiable through the process of mutilation in the proxy NCM (via Corol. 2).* ■



Corollary 4 (Soundness and Completeness). *Let Ω^* be the set of all SCMs, $\mathcal{M}^* \in \Omega^*$ be the true SCM inducing causal diagram $\mathcal{G}$, $Q = P(\mathbf{y} \mid do(\mathbf{x}))$ be a query of interest, and $\widehat{Q}$ 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 $\widehat{Q} = P^{\mathcal{M}^*}(\mathbf{y} \mid do(\mathbf{x}))$.* ■

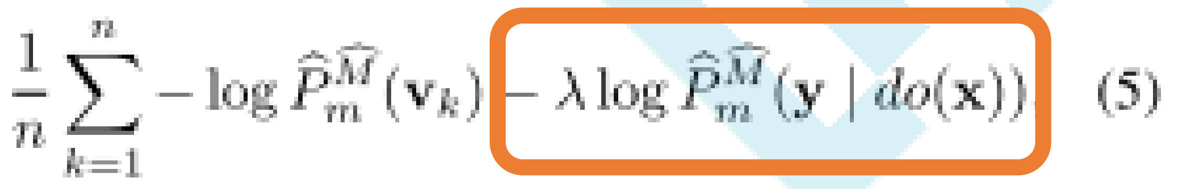
Algorithm



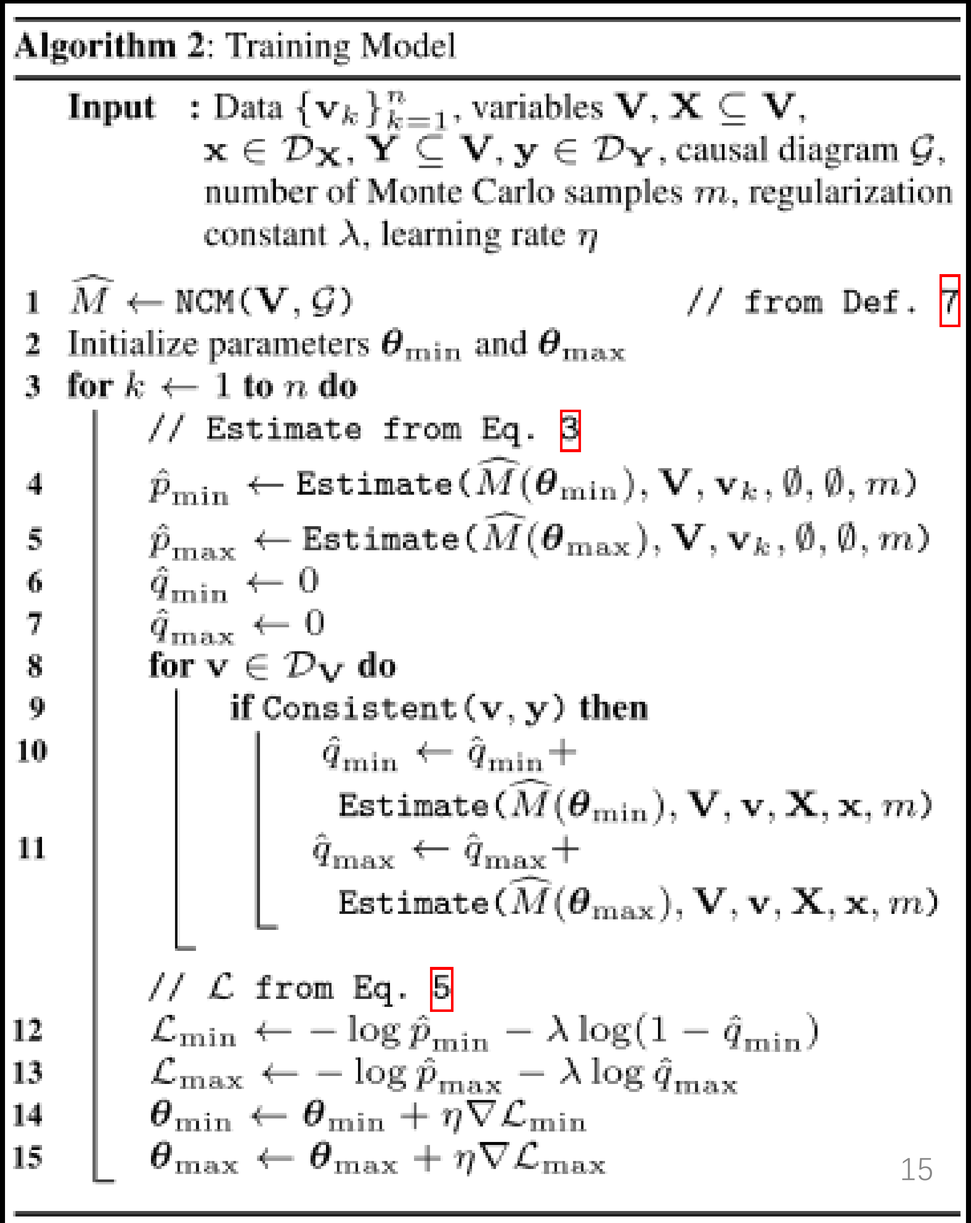
Algorithm 1: Identifying/estimating queries with NCMs.

```
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
```



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

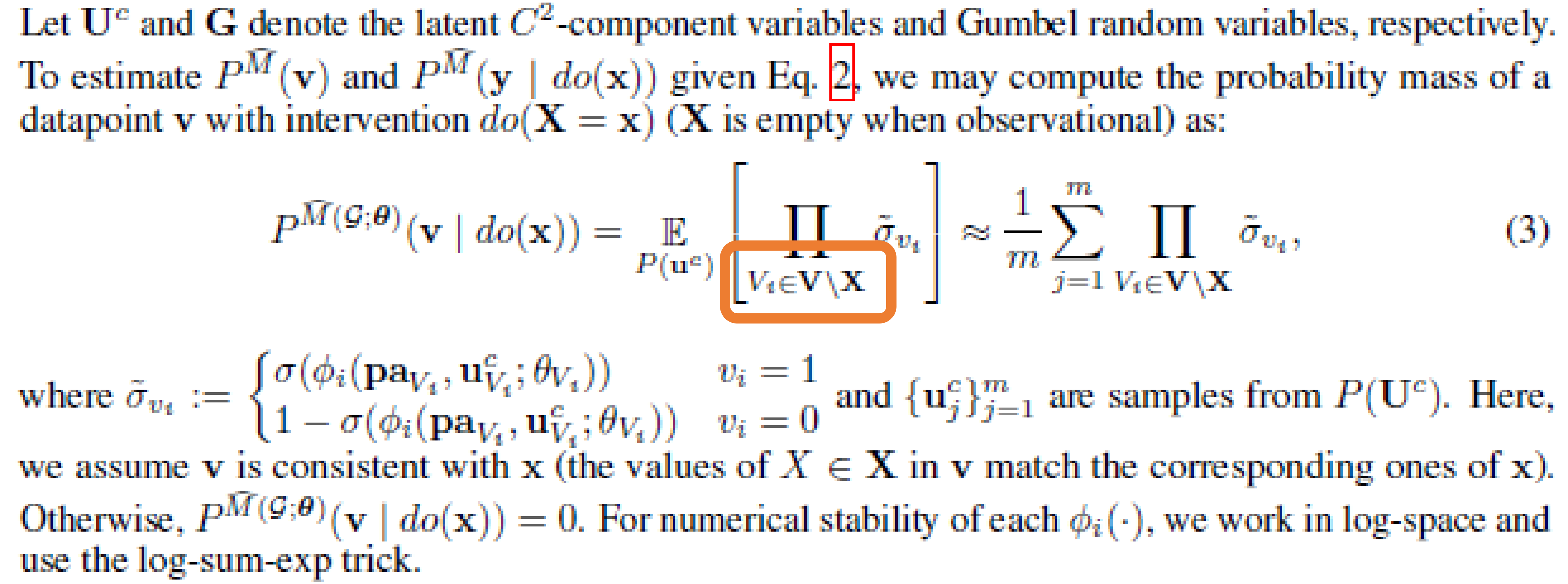


Algorithm 2: Training Model

```
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 $\lambda$, learning rate $\eta$

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

Algorithm

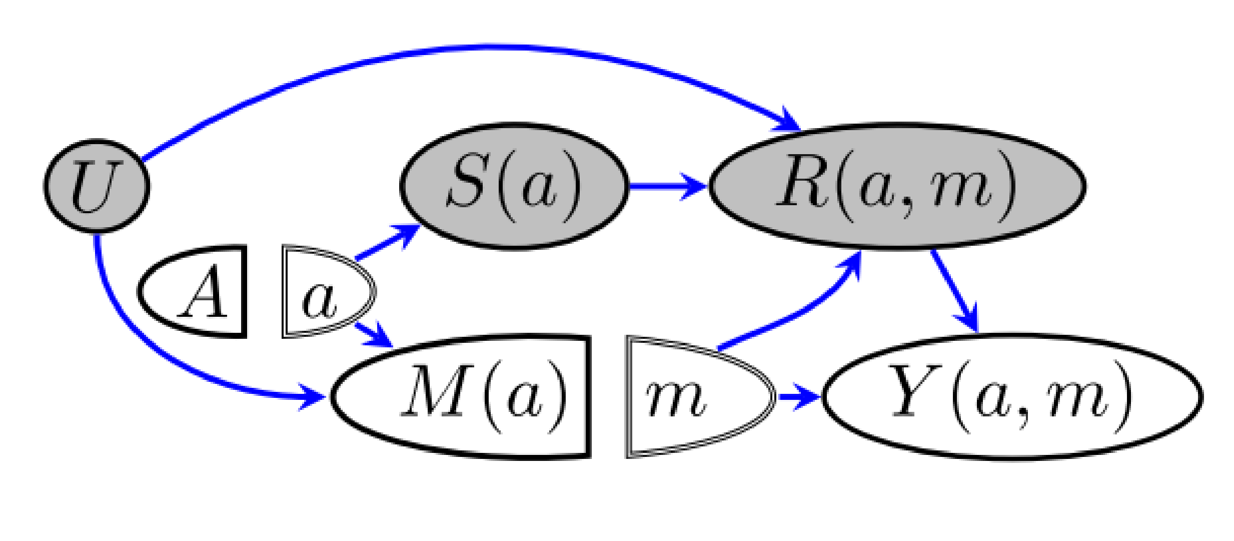


Let $\mathbf{U}^c$ and $\mathbf{G}$ denote the latent C^2-component variables and Gumbel random variables, respectively. To estimate $P^{\widehat{M}}(\mathbf{v})$ and $P^{\widehat{M}}(\mathbf{y} \mid do(\mathbf{x}))$ given Eq. $\boxed{2}$, we may compute the probability mass of a datapoint $\mathbf{v}$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

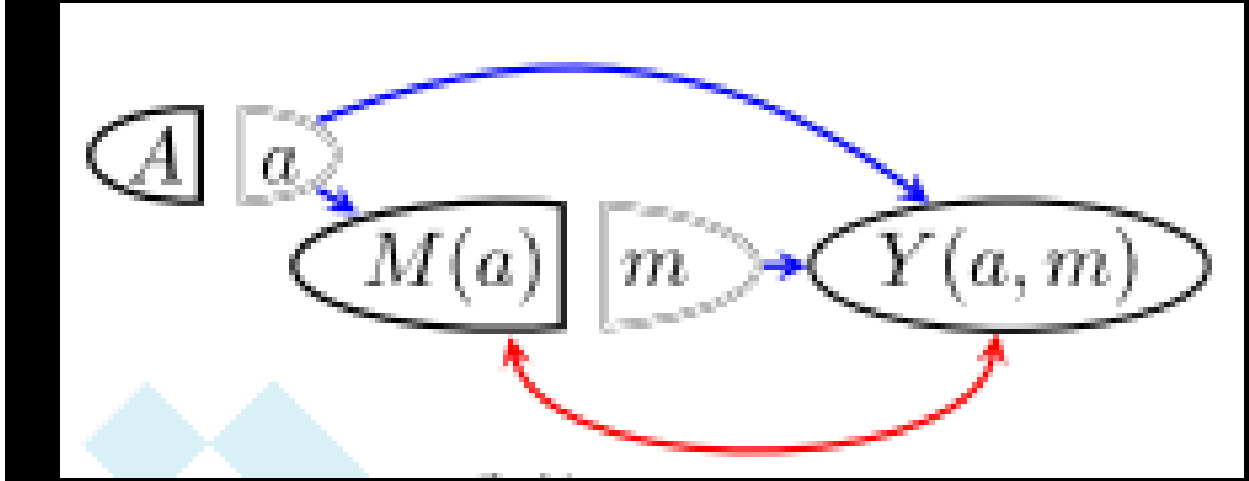
$$P^{\widehat{M}(\mathcal{G};\boldsymbol{\theta})}(\mathbf{v} \mid do(\mathbf{x})) = \mathop{\mathbb{E}}_{P(\mathbf{u}^c)}\left[\prod_{\boxed{V_i \in \mathbf{V}\setminus\mathbf{X}}} \tilde{\sigma}_{v_i}\right] \approx \frac{1}{m}\sum_{j=1}^{m} \prod_{V_i \in \mathbf{V}\setminus\mathbf{X}} \tilde{\sigma}_{v_i},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

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.

Example of mutilation/truncated factorization



SWIG (single world intervention graph)



SWIG without latent projection

Shpitser, I., Richardson, T.S. and **Robins**, J.M., 2020. Multivariate counterfactual systems and causal graphical models. *arXiv preprint arXiv:2008.06017*.

Algorithm

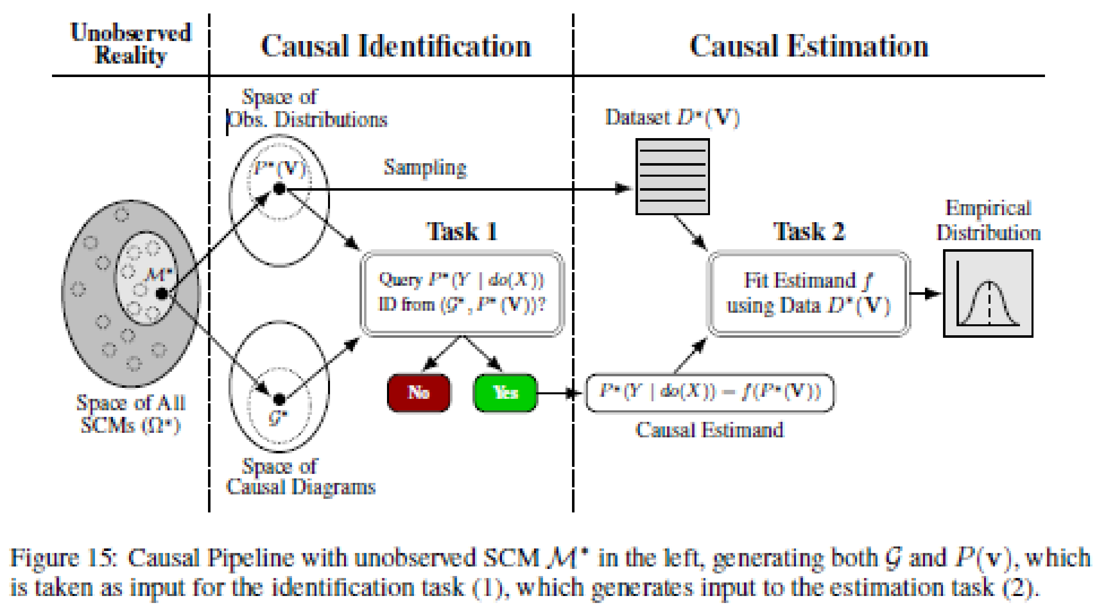
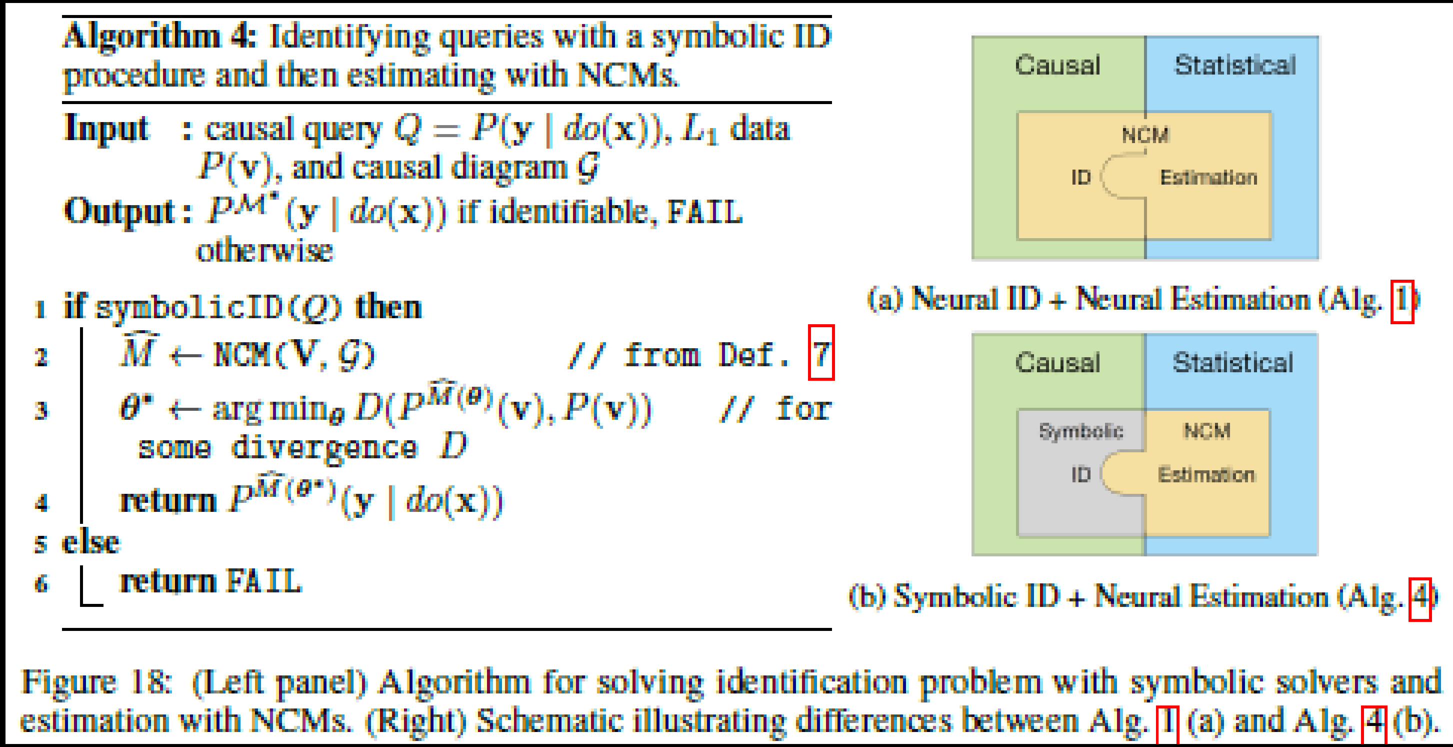

Unobserved Reality
Causal Identification
Causal Estimation
Space of Obs. Distributions
$P^*(\mathbf{V})$
Sampling
Dataset $D^*(\mathbf{V})$
Task 1
Task 2
Empirical Distribution
$\mathcal{M}^*$
Query $P^*(Y \mid do(X))$ ID from $(\mathcal{G}^*, P^*(\mathbf{V}))$?
Fit Estimand f using Data $D^*(\mathbf{V})$
No
Yes
$P^*(Y \mid do(X)) = f(P^*(\mathbf{V}))$
Causal Estimand
Space of All SCMs (Ω^*)
$\mathcal{G}^*$
Space of Causal Diagrams

Figure 15: Causal Pipeline with unobserved SCM $\mathcal{M}^*$ in the left, generating both $\mathcal{G}$ and $P(\mathbf{v})$, which is taken as input for the identification task (1), which generates input to the estimation task (2).



Algorithm 4: Identifying queries with a symbolic ID procedure and then estimating 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

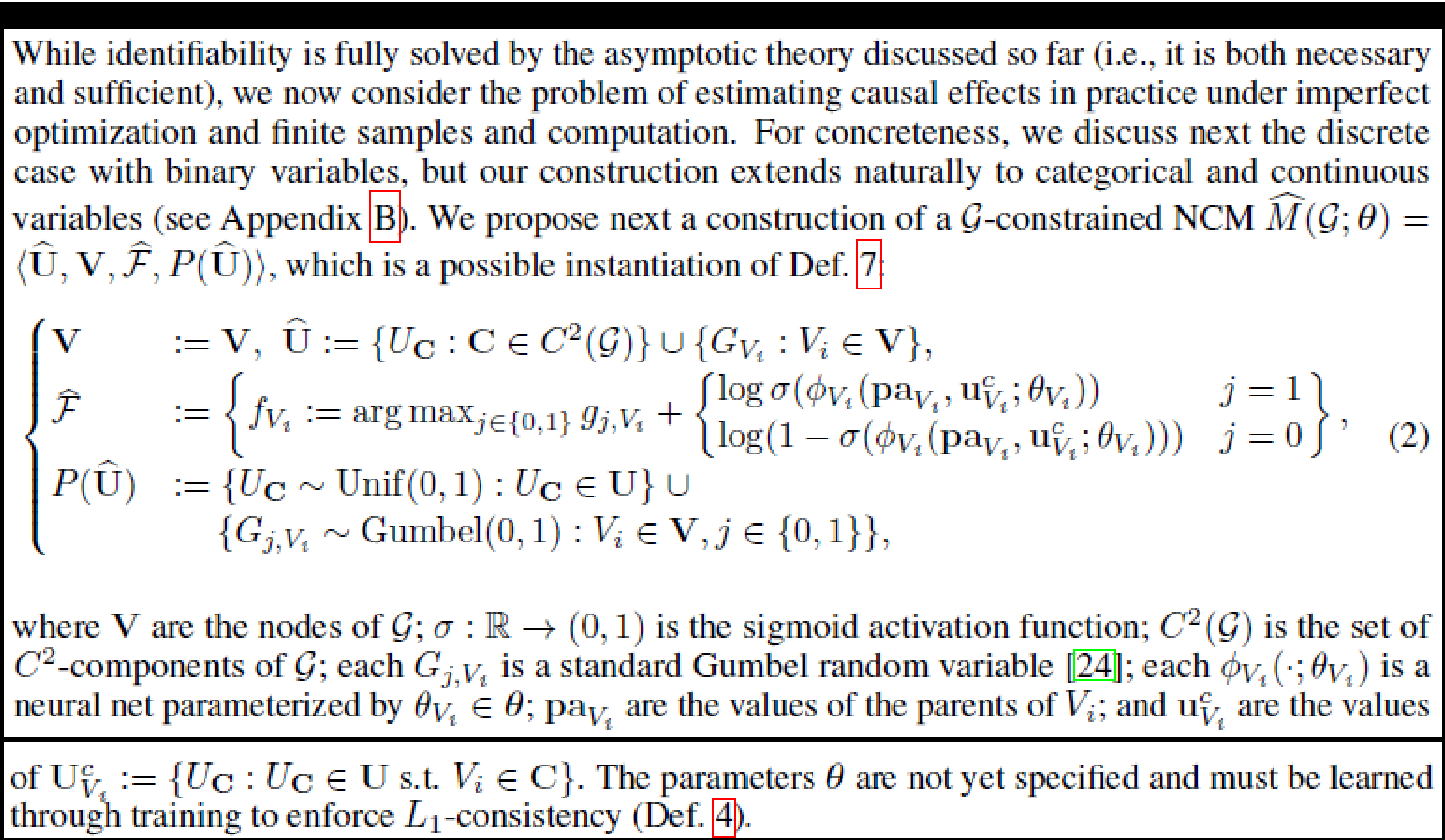
```
if symbolicID($Q$) then
    $\widehat{M} \leftarrow$ NCM($\mathbf{V}, \mathcal{G}$)                // from Def. 7
    $\theta^* \leftarrow \arg\min_{\theta} D(P^{\widehat{M}(\theta)}(\mathbf{v}), P(\mathbf{v}))$    // for
      some divergence $D$
    return $P^{\widehat{M}(\theta^*)}(\mathbf{y} \mid do(\mathbf{x}))$
else
    return FAIL
```

(a) Neural ID + Neural Estimation (Alg. 1)

(b) Symbolic ID + Neural Estimation (Alg. 4)

Figure 18: (Left panel) Algorithm for solving identification problem with symbolic solvers and estimation with NCMs. (Right) Schematic illustrating differences between Alg. 1 (a) and Alg. 4 (b).

Experiment Setting



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

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

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 [24]; each $\phi_{V_i}(\cdot;\theta_{V_i})$ is a neural net parameterized by $\theta_{V_i} \in \theta$; $\mathbf{pa}_{V_i}$ 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 θ are not yet specified and must be learned through training to enforce L_1-consistency (Def. 4).

Result

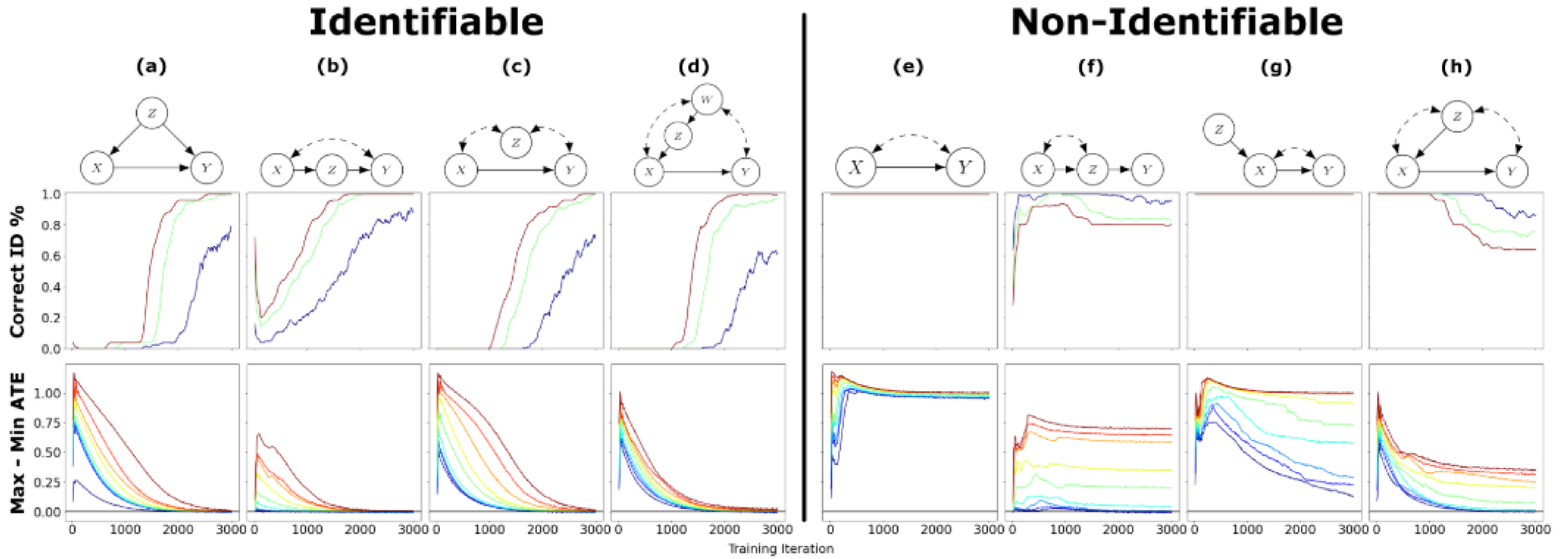


Figure 4: Experimental results on deciding identifiability with NCMs. **Top**: Graphs from left to right: (ID cases) back-door, front-door, M, napkin; (not ID cases) bow, extended bow, IV, bad M. **Middle**: Classification accuracy over 3,000 training epochs from running hypothesis test on Eq. 6 with $\tau = 0.01$ (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.

Result

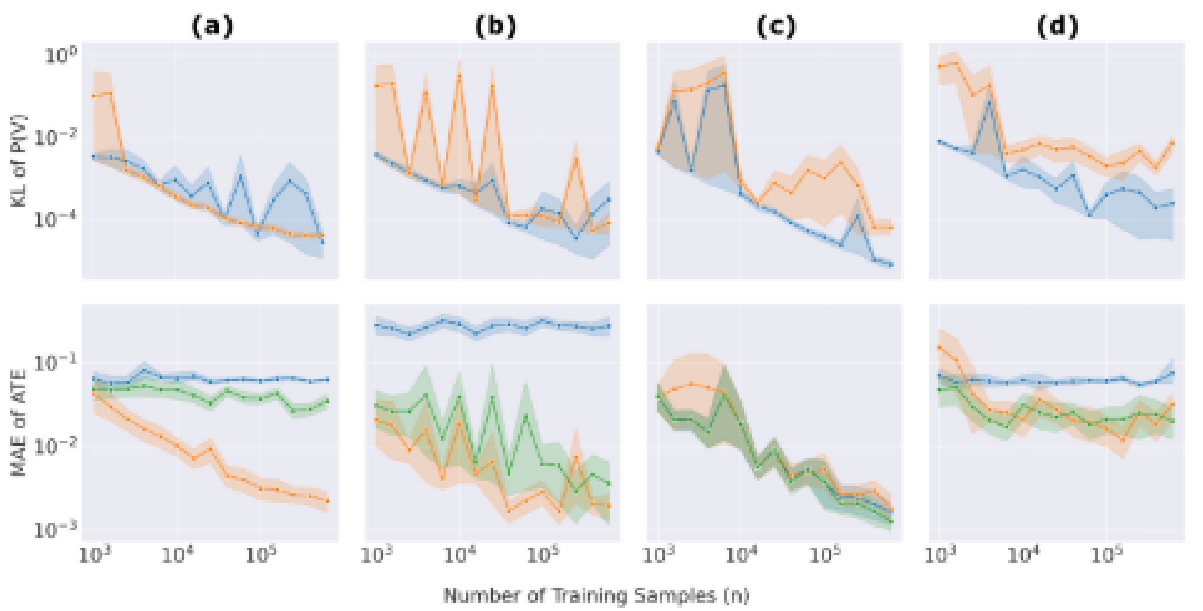


Figure 5: NCM estimation results for ID cases. Columns a, b, c, d correspond to the same graphs as a, b, c, d in Fig. 4. **Top**: KL divergence of $P(\mathbf{V})$ induced by naïve model (blue) and NCM (orange) compared to $P^{\mathcal{M}^*}(\mathbf{V})$. **Bottom**: MAE of ATE of naïve model (blue),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

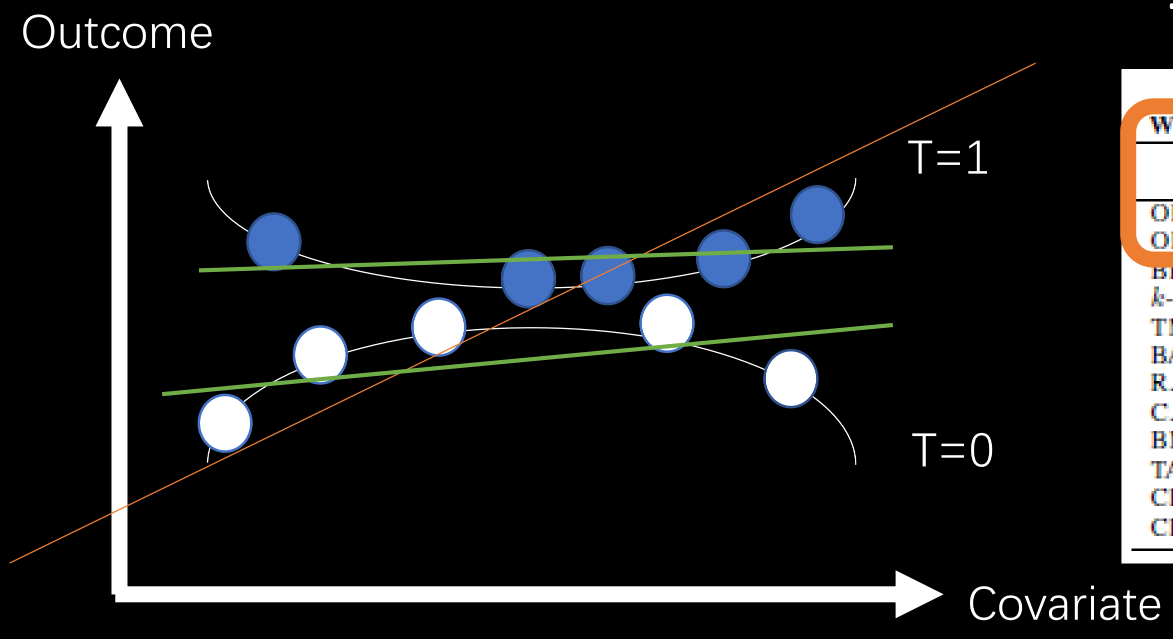
Paper 2: Interventional Sum-Product Networks: Causal Inference with Tractable Probabilistic Models

- Motivation: tractable causal models, solve intractability and flexibility challenge
- Methods: learning interventional distribution using CSPNs
- Experimental dataset (four toys): ASIA, Earthquake, Cancer, Causal Health
- Major findings: connect causality with tractable probabilistic models by using SPNs

Why learn interventional pdf rather than observational pdf ?

ignoring causal change(s) in a system, i.e., the change of structural equation(s) underlying the system, can lead to a significant performance decrease and safety hazards.

---- *Elements of causal inference*



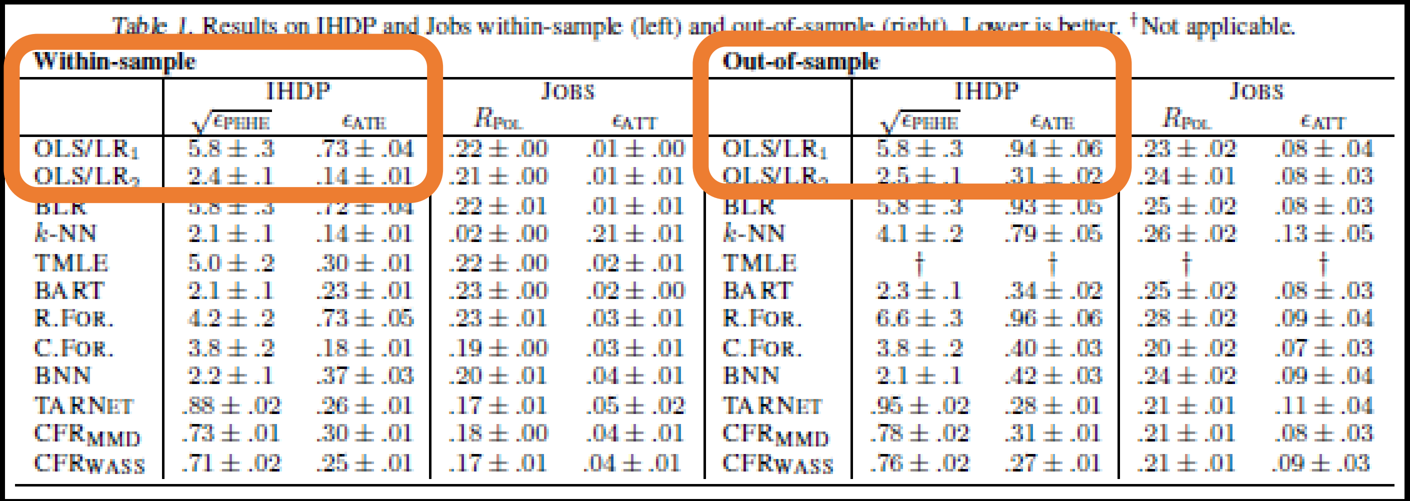


Table 1. Results on IHDP and Jobs within-sample (left) and out-of-sample (right). Lower is better. †Not applicable.

Within-sample	IHDP		Jobs		Out-of-sample	IHDP		Jobs	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	R_{POL}	ϵ_{ATT}		$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	R_{POL}	ϵ_{ATT}
OLS/LR$_1$	5.8 ± .3	.73 ± .04	.22 ± .00	.01 ± .00	OLS/LR$_1$	5.8 ± .3	.94 ± .06	.23 ± .02	.08 ± .04
OLS/LR$_2$	2.4 ± .1	.14 ± .01	.21 ± .00	.01 ± .01	OLS/LR$_2$	2.5 ± .1	.31 ± .02	.24 ± .01	.08 ± .03
BLR	5.8 ± .3	.72 ± .04	.22 ± .01	.01 ± .01	BLR	5.8 ± .3	.93 ± .05	.25 ± .02	.08 ± .03
k-NN	2.1 ± .1	.14 ± .01	.02 ± .00	.21 ± .01	k-NN	4.1 ± .2	.79 ± .05	.26 ± .02	.13 ± .05
TMLE	5.0 ± .2	.30 ± .01	.22 ± .00	.02 ± .01	TMLE	†	†	†	†
BART	2.1 ± .1	.23 ± .01	.23 ± .00	.02 ± .00	BART	2.3 ± .1	.34 ± .02	.25 ± .02	.08 ± .03
R.For.	4.2 ± .2	.73 ± .05	.23 ± .01	.03 ± .01	R.For.	6.6 ± .3	.96 ± .06	.28 ± .02	.09 ± .04
C.For.	3.8 ± .2	.18 ± .01	.19 ± .00	.03 ± .01	C.For.	3.8 ± .2	.40 ± .03	.20 ± .02	.07 ± .03
BNN	2.2 ± .1	.37 ± .03	.20 ± .01	.04 ± .01	BNN	2.1 ± .1	.42 ± .03	.24 ± .02	.09 ± .04
TARNet	.88 ± .02	.26 ± .01	.17 ± .01	.05 ± .02	TARNet	.95 ± .02	.28 ± .01	.21 ± .01	.11 ± .04
CFR$_{MMD}$	.73 ± .01	.30 ± .01	.18 ± .00	.04 ± .01	CFR$_{MMD}$	.78 ± .02	.31 ± .01	.21 ± .01	.08 ± .03
CFR$_{WASS}$	.71 ± .02	.25 ± .01	.17 ± .01	.04 ± .01	CFR$_{WASS}$	.76 ± .02	.27 ± .01	.21 ± .01	.09 ± .03

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference. The MIT Press, 2017.
Shalit, U., Johansson, F.D. and Sontag, D., 2017, July. Estimating individual treatment effect: generalization bounds and algorithms. In International Conference on Machine Learning (pp. 3076-3085). PMLR.

What's SPN?

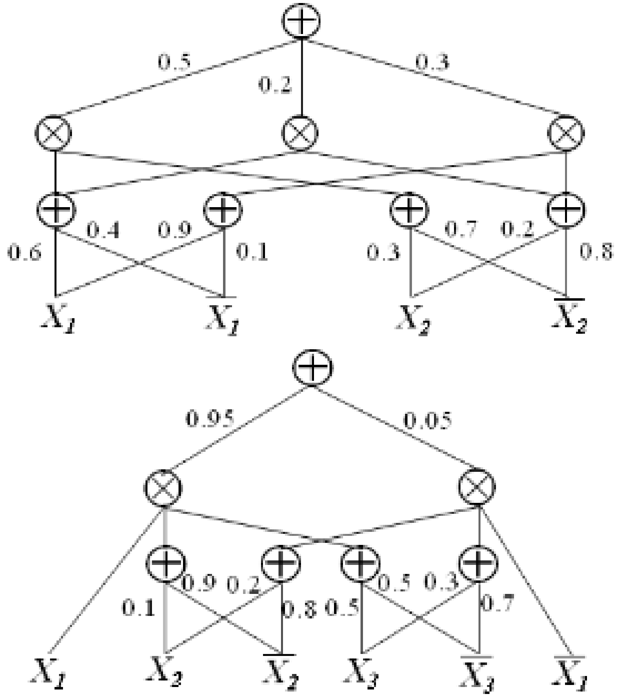


Figure 1. Left: SPN implementing a naive Bayes mixture model (three components,two variables). Right: SPN implementing a junction tree (clusters (X_1, X_2) and (X_1, X_3), separator X_1).

Advantage

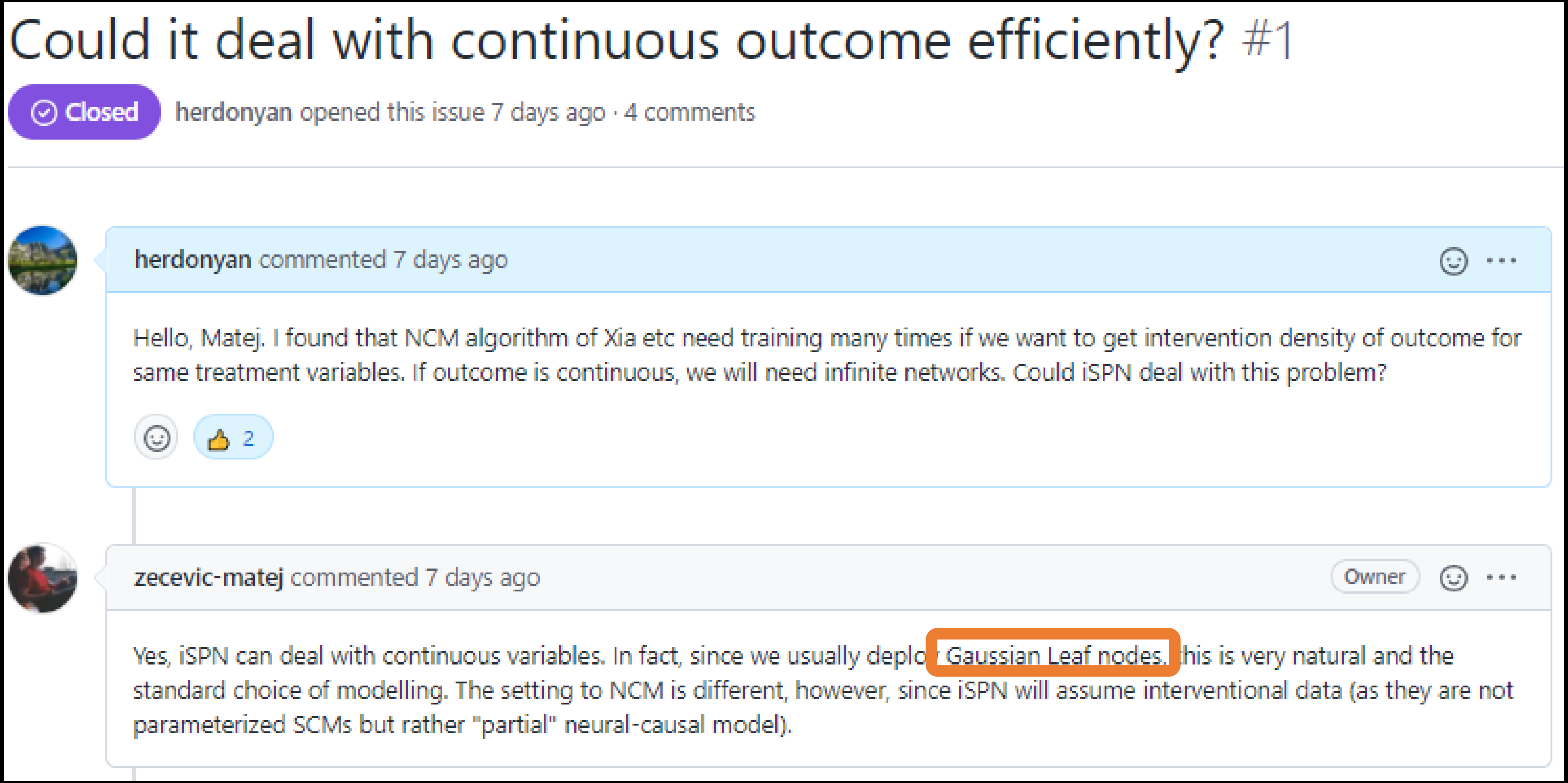
- Inference tasks in time proportional to the number of links in the graph
- Compute ANY joint, marginal, or conditional probability at most two upward passes with one network

H. Poon and P. Domingos, "Sum-product networks: A new deep architecture," 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011, pp. 689-690, doi: 10.1109/ICCVW.2011.6130310.

Gens, R. and Domingos, P., 2012. Discriminative learning of sum-product networks. Advances in Neural Information Processing Systems, 25, pp.3239-3247. (Outstanding Student Paper Awards)

Sánchez-Cauce, R., París, I. and Díez, F.J., 2021. Sum-product networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.

What’s SPN?



Could it deal with continuous outcome efficiently? #1

herdonyan opened this issue 7 days ago · 4 comments

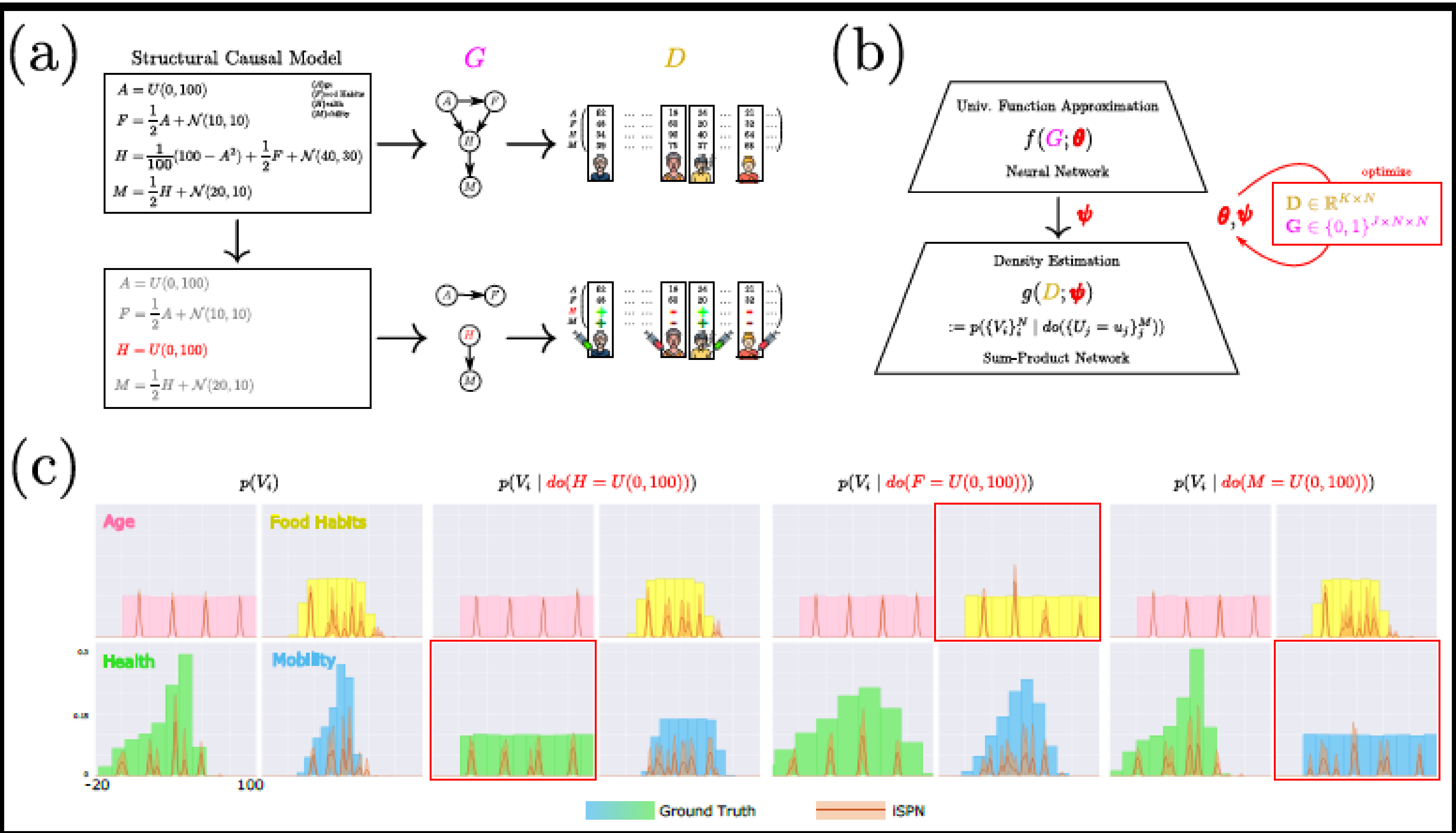
herdonyan commented 7 days ago

Hello, Matej. I found that NCM algorithm of Xia etc need training many times if we want to get intervention density of outcome for same treatment variables. If outcome is continuous, we will need infinite networks. Could iSPN deal with this problem?

zecevic-matej commented 7 days ago

Yes, iSPN can deal with continuous variables. In fact, since we usually deploy Gaussian Leaf nodes, this is very natural and the standard choice of modelling. The setting to NCM is different, however, since iSPN will assume interventional data (as they are not parameterized SCMs but rather "partial" neural-causal model).

An overview of iSPN


(a)
Structural Causal Model
$A = U(0, 100)$
$F = \frac{1}{2}A + \mathcal{N}(10, 10)$
$H = \frac{1}{100}(100 - A^2) + \frac{1}{2}F + \mathcal{N}(40, 30)$
$M = \frac{1}{2}H + \mathcal{N}(20, 10)$
(A)ge
(F)ood Habits
(H)ealth
(M)obility
G
D
A
F
H
M
$A = U(0, 100)$
$F = \frac{1}{2}A + \mathcal{N}(10, 10)$
$H = U(0, 100)$
$M = \frac{1}{2}H + \mathcal{N}(20, 10)$
(b)
Univ. Function Approximation
$f(G; \boldsymbol{\theta})$
Neural Network
$\boldsymbol{\psi}$
Density Estimation
$g(D; \boldsymbol{\psi})$
$:= p(\{V_i\}_i^N \mid do(\{U_j = u_j\}_j^M))$
Sum-Product Network
$\boldsymbol{\theta}, \boldsymbol{\psi}$
optimize
$\mathbf{D} \in \mathbb{R}^{K \times N}$
$\mathbf{G} \in \{0, 1\}^{J \times N \times N}$
(c)
$p(V_i)$
$p(V_i \mid do(H = U(0, 100)))$
$p(V_i \mid do(F = U(0, 100)))$
$p(V_i \mid do(M = U(0, 100)))$
Age
Food Habits
Health
Mobility
0.3
0.15
0
-20
100
Ground Truth
iSPN

Xiaoting Shao, Alejandro Molina, Antonio Vergari, Karl Stelzner, Robert Peharz, Thomas Liebig, and Kristian Kersting. Conditional sum-product networks: Imposing structure on deep probabilistic architectures. arXiv preprintarXiv:1905.08550, 2019.

An overview of iSPN

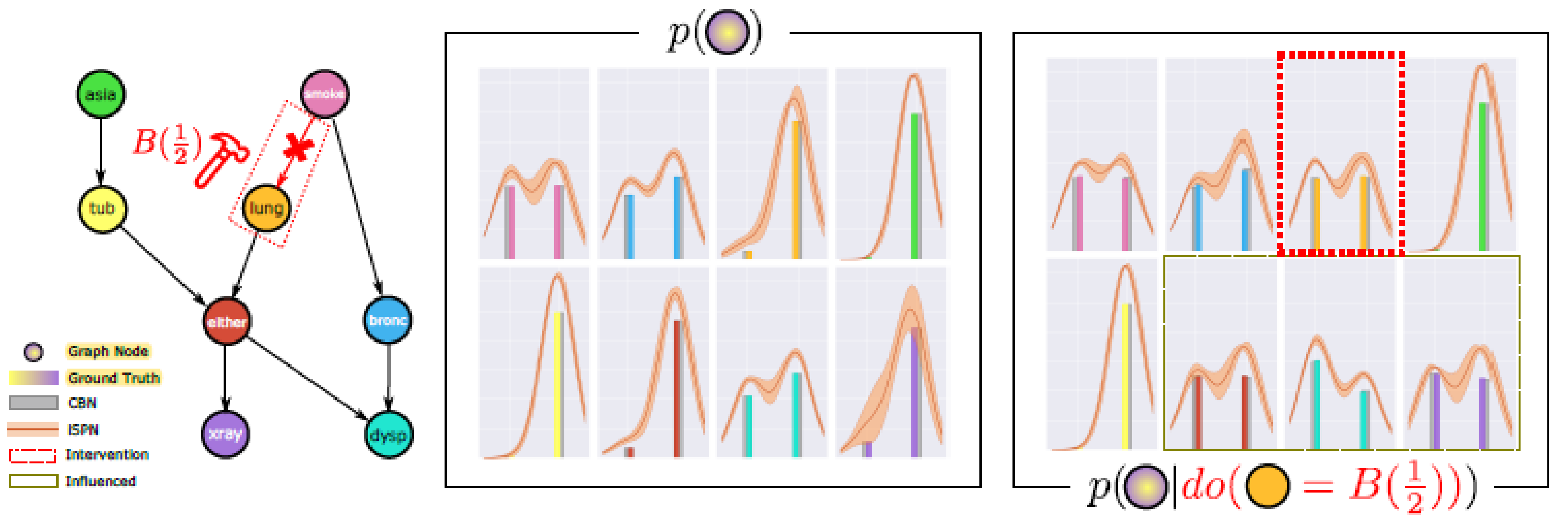
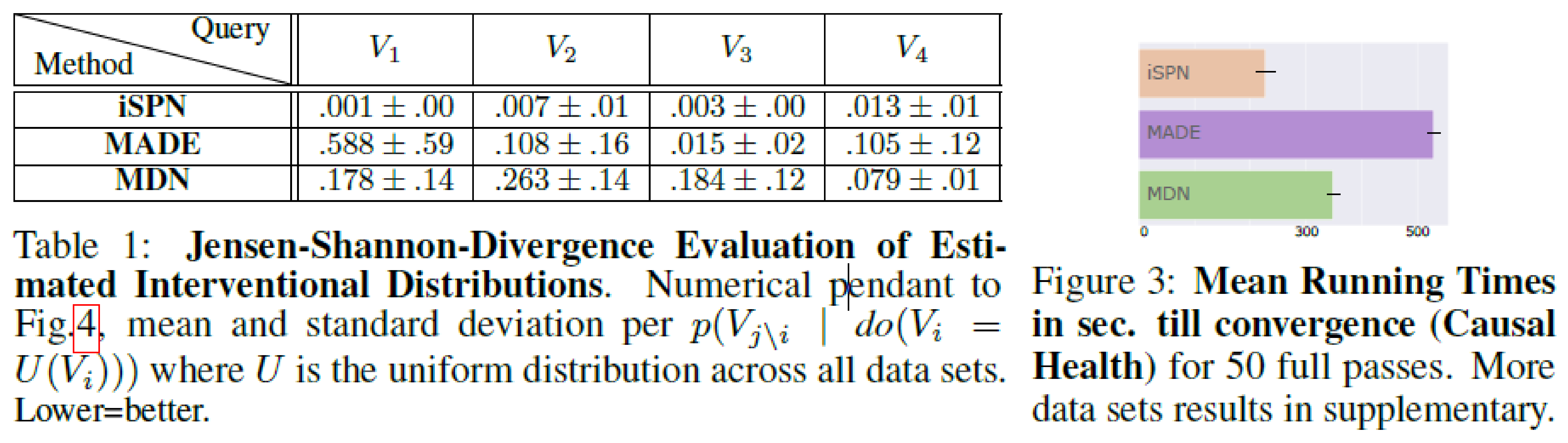


Figure 1: **Capturing interventional distributions using iSPN**. The interventional distributions for the ASIA data set using a causal Bayesian network (CBN, small-scale gold standard, gray bars) as well as an interventional SPN (iSPN) by intervening on *lung*. The iSPN is sensible to all the influences of the given intervention onto the system i.e., subsequent effects in the causal hierachy. (Best viewed in color.)

Experiment

- Comparison to Generative models
- Comparison of running times



Query / Method	V_1	V_2	V_3	V_4
iSPN	.001 ± .00	.007 ± .01	.003 ± .00	.013 ± .01
MADE	.588 ± .59	.108 ± .16	.015 ± .02	.105 ± .12
MDN	.178 ± .14	.263 ± .14	.184 ± .12	.079 ± .01

Table 1: **Jensen-Shannon-Divergence Evaluation of Estimated Interventional Distributions.** Numerical pendant to Fig.4, mean and standard deviation per $p(V_{j\setminus i} \mid do(V_i = U(V_i)))$ where U is the uniform distribution across all data sets. Lower=better.

Figure 3: **Mean Running Times in sec. till convergence (Causal Health)** for 50 full passes. More data sets results in supplementary.

Experiment

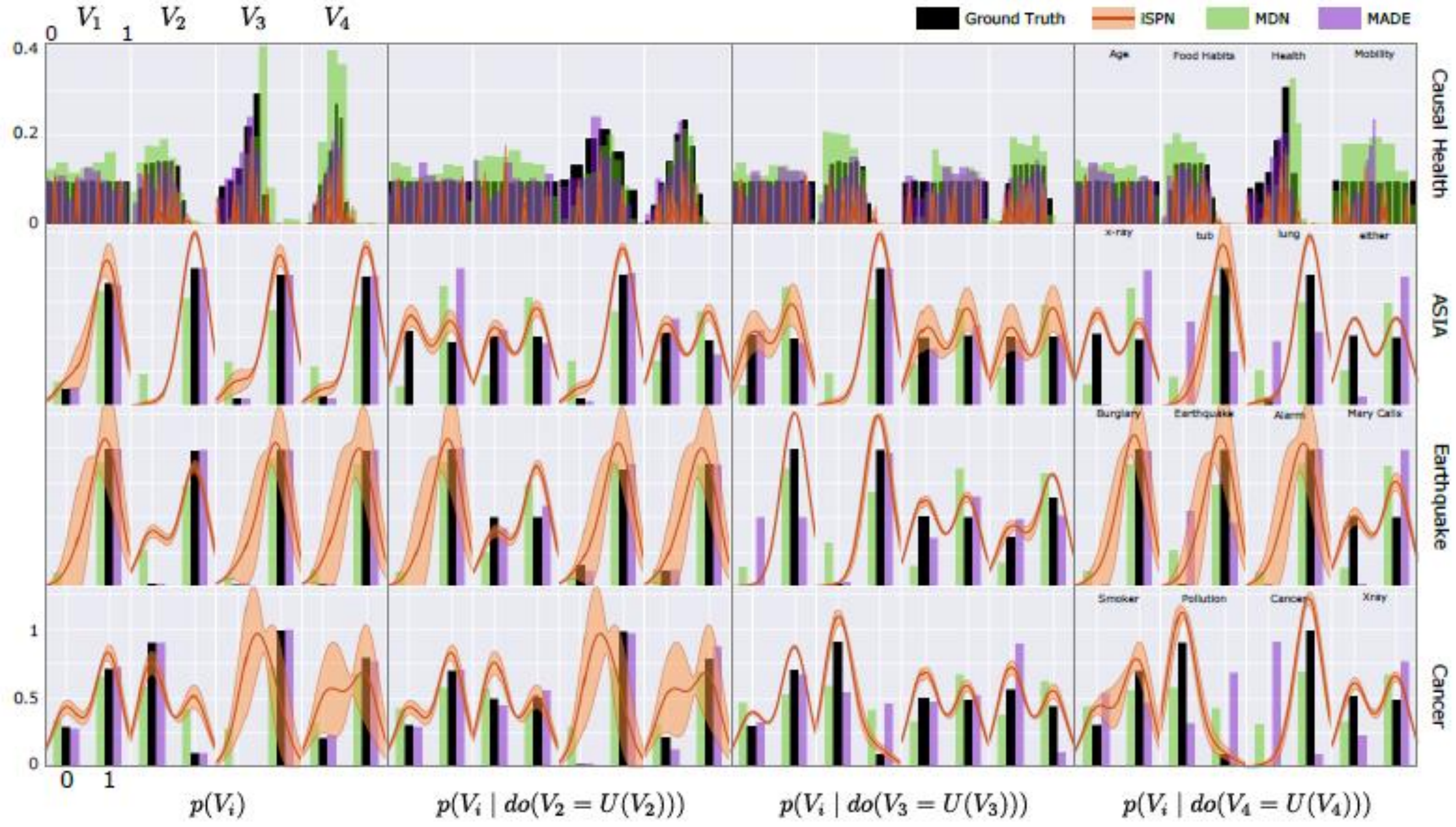
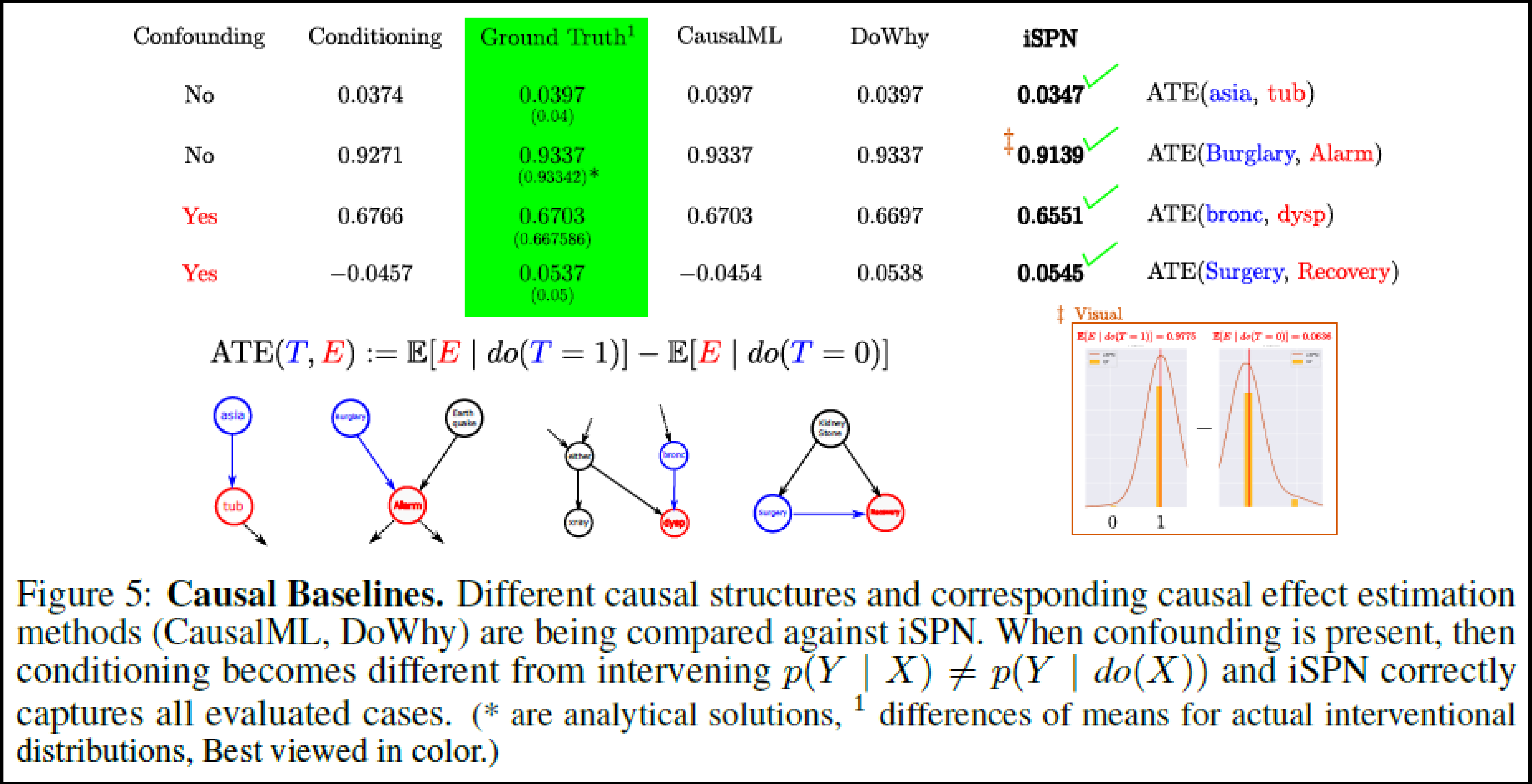


Figure 4: **Generative Baselines.** A comparison to the ground-truth (via underlying SCM) and competing estimated distributions. Each row represents a data set and each column represents a variable for a given causal query. (Best in color.)

- **Precise Estimation of do-influenced variables**
- **Comparison to Generative models**

Experiment



Confounding	Conditioning	Ground Truth[1]	CausalML	DoWhy	iSPN	
No	0.0374	0.0397 (0.04)	0.0397	0.0397	**0.0347** ✓	ATE(asia, tub)
No	0.9271	0.9337 (0.93342)*	0.9337	0.9337	‡**0.9139** ✓	ATE(Burglary, Alarm)
Yes	0.6766	0.6703 (0.667586)	0.6703	0.6697	**0.6551** ✓	ATE(bronc, dysp)
Yes	−0.0457	0.0537 (0.05)	−0.0454	0.0538	**0.0545** ✓	ATE(Surgery, Recovery)

$$\mathrm{ATE}(T, E) := \mathbb{E}[E \mid do(T=1)] - \mathbb{E}[E \mid do(T=0)]$$

asia
tub
Burglary
Earth quake
Alarm
either
xray
bronc
dysp
Kidney Stone
Surgery
Recovery
‡ Visual
$\mathbb{E}[E \mid do(T=1)] = 0.9775$
$\mathbb{E}[E \mid do(T=0)] = 0.0686$
−
0
1

Figure 5: **Causal Baselines.** Different causal structures and corresponding causal effect estimation methods (CausalML, DoWhy) are being compared against iSPN. When confounding is present, then conditioning becomes different from intervening $p(Y \mid X) \neq p(Y \mid do(X))$ and iSPN correctly captures all evaluated cases. (* are analytical solutions, [1] differences of means for actual interventional distributions, Best viewed in color.)

- **Comparison to causal models**

Experiment

■ Different Types of Intervention

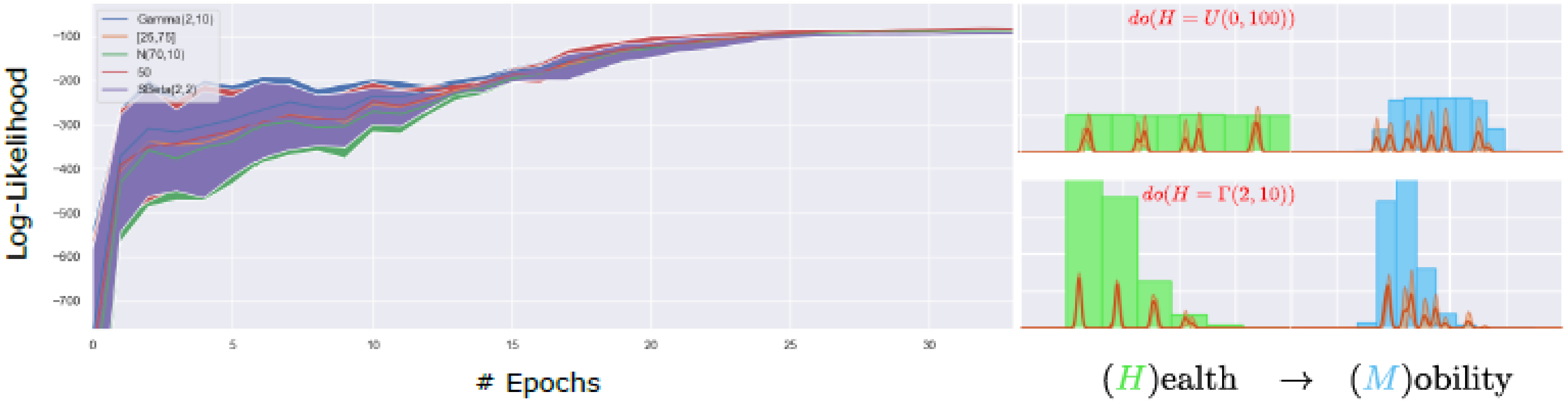


Figure 6: **Adaptation to Different Interventions.** Training results for different kinds of interventions on the continuous CH data set. Left, the respective mean objective curves (log-likelihood), indicating consistent training and convergence for all three random seeds per configuration. Right, the (mean) density functions for two different interventions on H: Uniform $U(a, b)$ and Gamma $\Gamma(p, q)$ (other interventions shown in the supplementary). (Best viewed in color.)

Limitations and further enhancements

	Data	Query	Response	Tractable	Explicit Inductive Bias
ID-NCM (paper 1)	L_1	L_2	L_0	✗	✓
iSPN (paper 2)	L_2	L_2	L_0	✓	✓
Reinforcement learning	L_1 / L_2 / L_3	L_2	L_0 / L_1 / L_2	✓	✗

Limitations and further enhancements

Limitation:

require **DAG** hypothesis; **high dimensional** problem; lack **mechanism explanation**; neural inductive bias; **toy** experiments

Paper 1: not tractable (training one network only for one specific intervention query)

Paper 2: learning from intervention data

Further enhancements:

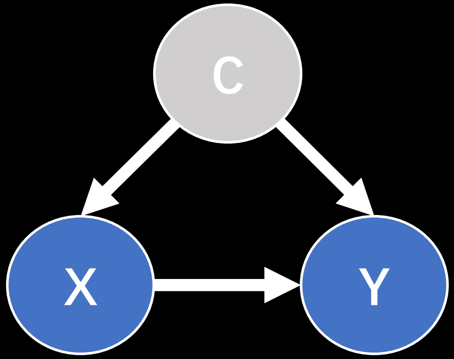
address **realistic and large dataset**, such as MIMIC-IV and MIMIC-CXR; solve these problem using deep learning method

make individual treatment in reality be a killer-app of causal inference

Thanks for watching!

Paper 3: Non-Parametric Methods for Partial Identification of Causal Effects

- Motivation: give a tight bound of non-identifiable causal effects given data and generation hypothesis (DAG)
- Methods: causal diagram -> canonical diagram
- Experimental dataset: International Stroke Trial (1997)
- Major findings: link between causal diagram and canonical diagram; an efficient algorithm for bounding causal effects from observation in arbitrary causal diagrams



$P(Y \mid do(x))$ is non-identifiable.

Method

- Theorem 2. For a causal diagram G and its canonical diagram H, consider the following conditions:

1. M is the set of all SCMs associated with G.

2. N is the set of all SCMs associated with H where each Ri 2 R is a discrete variable drawn from N.

Then M and N are do-equivalent.

Method

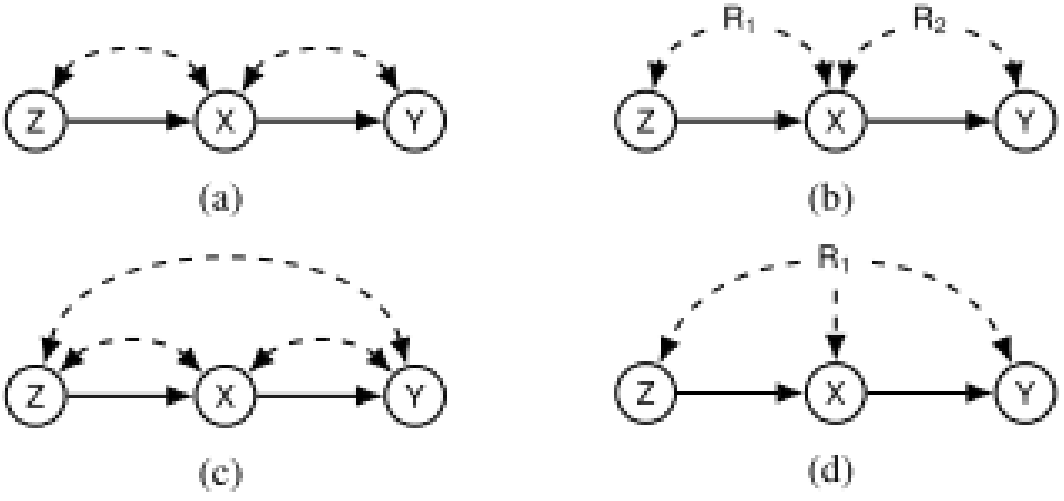


Figure 3: (a, c) causal diagrams $\mathcal{G}$; (b, d) canonical diagrams $\mathcal{H}$ where exogenous variables $\boldsymbol{R}$ are explicitly shown.

Algorithm 1 INVERSEPROJECT

```
Input: $\mathcal{G}$ and $\{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_K\}$ where $\boldsymbol{C}_k \subseteq \boldsymbol{V}$.
Output: A canonical diagram $\mathcal{H}$ where all exogenous
    variables $\boldsymbol{R}$ are shown explicitly.
For each node $V \in \mathcal{G}$, add a node $V$ in $\mathcal{H}$.
For each arrow $V_i \to V_j \in \mathcal{G}$, add $V_i \to V_j$ in $\mathcal{H}$.
For each $\boldsymbol{C}_k$, add an empty node $R_i$ in $\mathcal{H}$.
For each $V \in \boldsymbol{C}_k$, add an arrow $R_i \to V$.
```

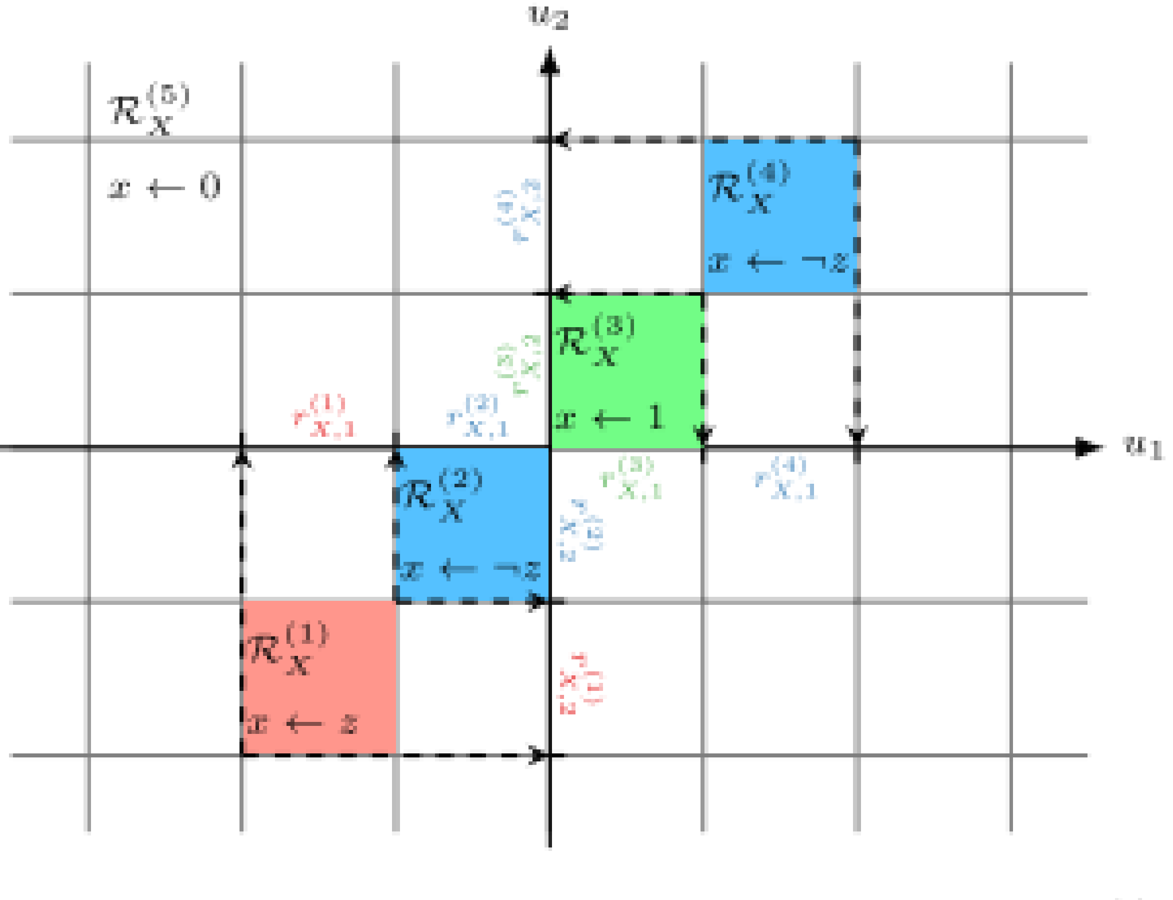
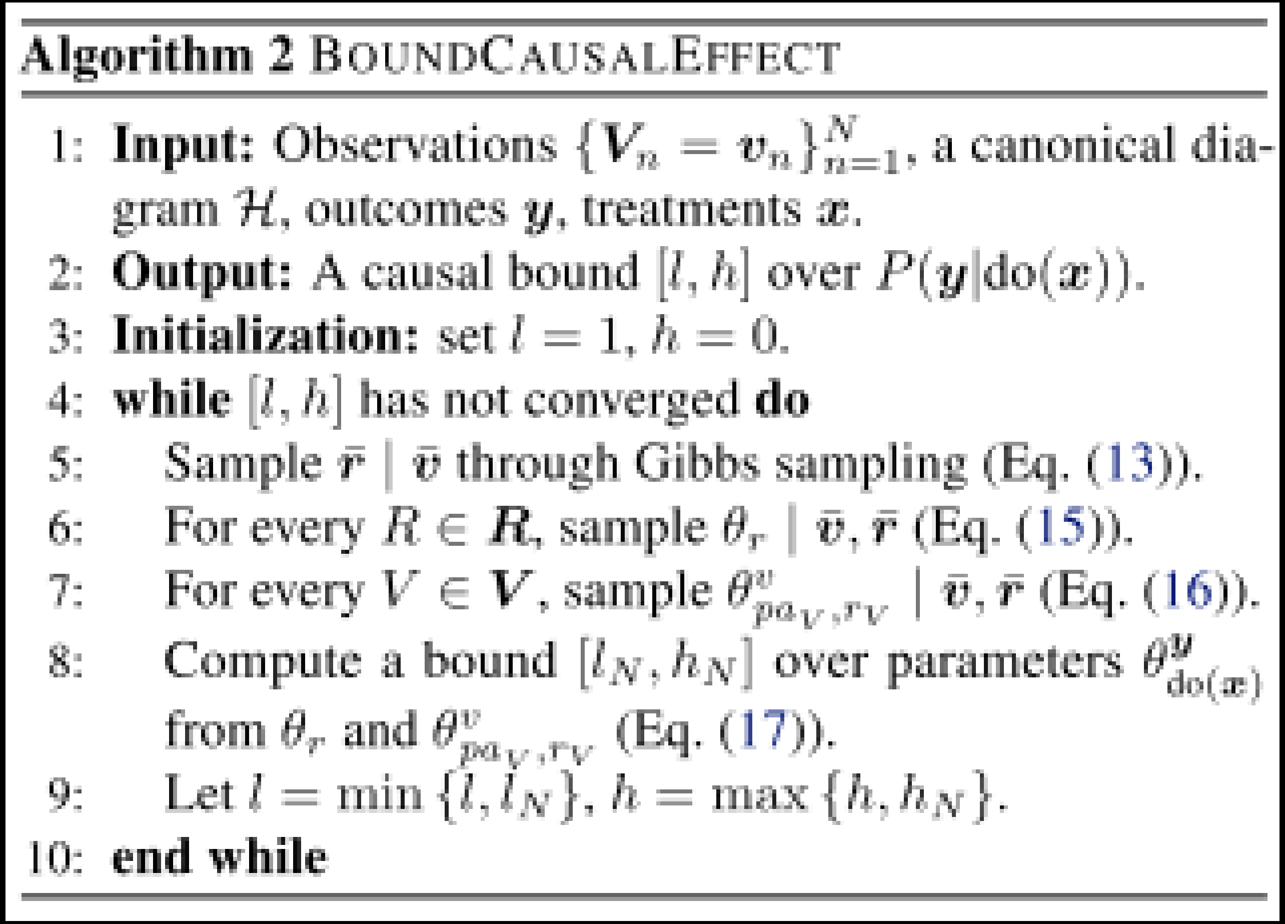


Figure 4: A graphical representation of rectangles $\mathcal{R}_X^{(i)}$.

Method

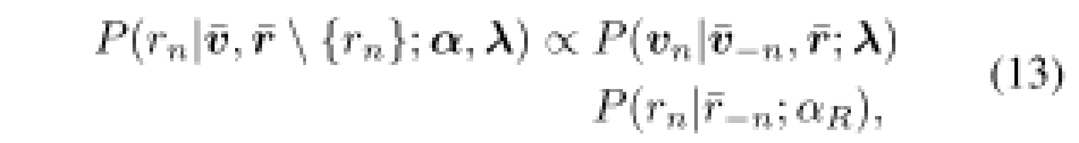
- $P(\mathrm{y}|do(x)) = \sum_{r,z} \prod_{V \in \boldsymbol{V} \setminus \boldsymbol{X}} P(v|pa_V, r_V) \prod_{R \in \boldsymbol{R}} P(r)$

Algorithm

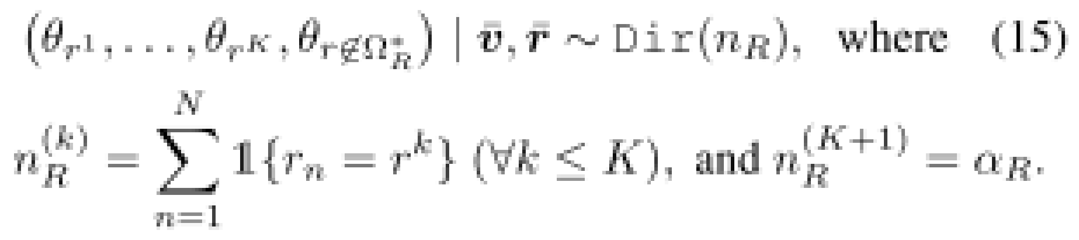


Algorithm 2 BoundCausalEffect

```
Input: Observations $\{\boldsymbol{V}_n = \boldsymbol{v}_n\}_{n=1}^N$, a canonical diagram $\mathcal{H}$, outcomes $\boldsymbol{y}$, treatments $\boldsymbol{x}$.
Output: A causal bound $[l, h]$ over $P(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$.
Initialization: set $l = 1$, $h = 0$.
while $[l, h]$ has not converged do
    Sample $\bar{\boldsymbol{r}} \mid \bar{\boldsymbol{v}}$ through Gibbs sampling (Eq. (13)).
    For every $R \in \boldsymbol{R}$, sample $\theta_r \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{r}}$ (Eq. (15)).
    For every $V \in \boldsymbol{V}$, sample $\theta^v_{pa_V, r_V} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{r}}$ (Eq. (16)).
    Compute a bound $[l_N, h_N]$ over parameters $\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x})}$ from $\theta_r$ and $\theta^v_{pa_V, r_V}$ (Eq. (17)).
    Let $l = \min\{l, l_N\}$, $h = \max\{h, h_N\}$.
end while
```

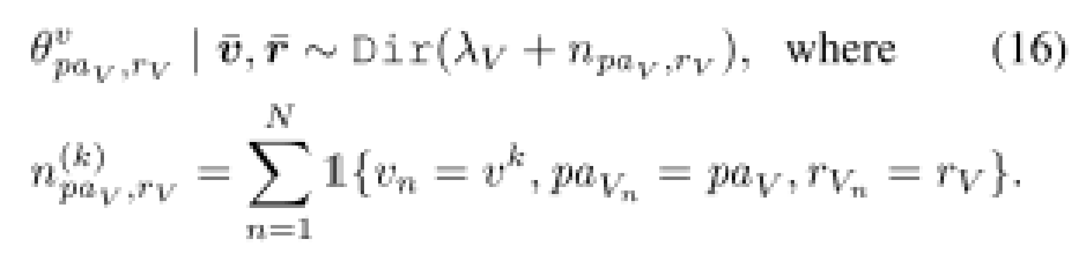


$$P(r_n|\bar{\boldsymbol{v}}, \bar{\boldsymbol{r}} \setminus \{r_n\}; \boldsymbol{\alpha}, \boldsymbol{\lambda}) \propto P(\boldsymbol{v}_n|\bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{r}}; \boldsymbol{\lambda}) \, P(r_n|\bar{r}_{-n}; \alpha_R), \tag{13}$$



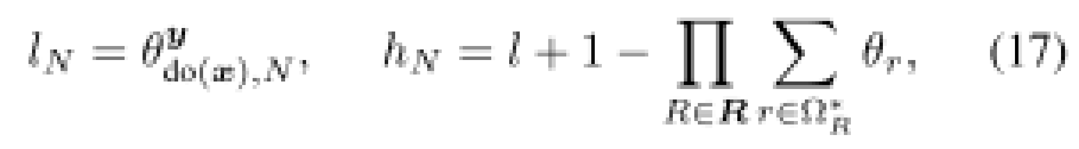
$$(\theta_{r^1}, \dots, \theta_{r^K}, \theta_{r \notin \Omega^*_R}) \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{r}} \sim \texttt{Dir}(n_R), \quad \text{where} \tag{15}$$

$$n_R^{(k)} = \sum_{n=1}^{N} \mathbb{1}\{r_n = r^k\} \; (\forall k \leq K), \text{ and } n_R^{(K+1)} = \alpha_R.$$



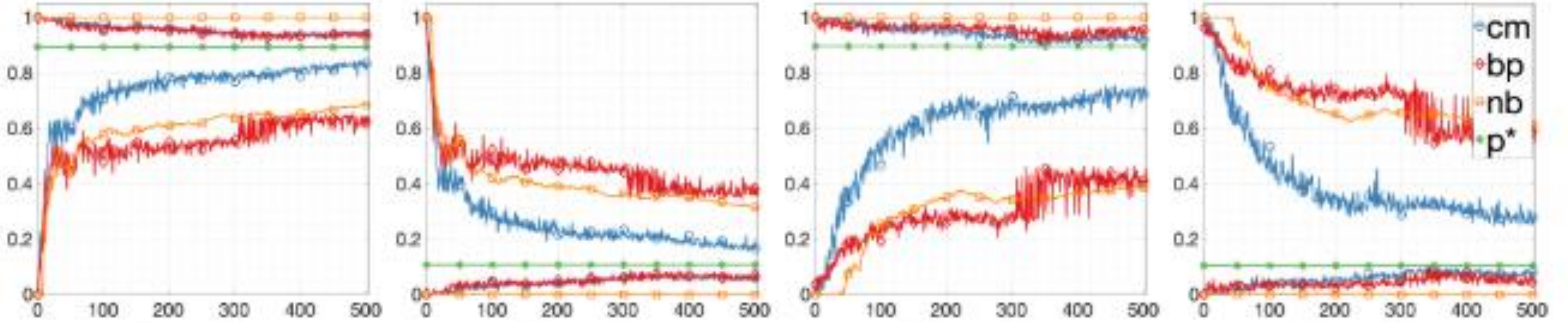
$$\theta^v_{pa_V, r_V} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{r}} \sim \texttt{Dir}(\lambda_V + n_{pa_V, r_V}), \quad \text{where} \tag{16}$$

$$n^{(k)}_{pa_V, r_V} = \sum_{n=1}^{N} \mathbb{1}\{v_n = v^k, pa_{V_n} = pa_V, r_{V_n} = r_V\}.$$



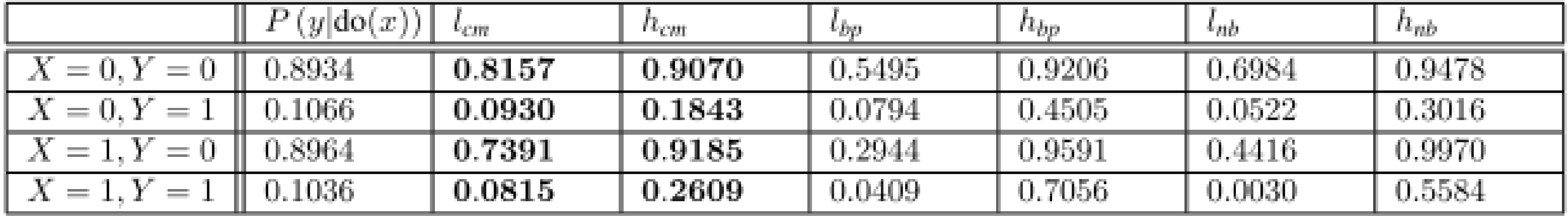
$$l_N = \theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x}), N}, \quad h_N = l + 1 - \prod_{R \in \boldsymbol{R}} \sum_{r \in \Omega^*_R} \theta_r, \tag{17}$$

Result



(a) $X = 0, Y = 0$ (b) $X = 0, Y = 1$ (c) $X = 1, Y = 0$ (d) $X = 1, Y = 1$

Figure 6: Causal bounds over the effect $P(y|\text{do}(x))$ of aspirin X on the death Y in the International Stroke Trial (IST). The x-axle represents the number of observational samples. *cm* are new bounds derived by Alg. 2 (blue); *bp* are derived using simple canonical partitions (red); *nb* are the natural bounds (yellow). The actual effect $P(y|do(x))$ is labeled as $p*$ (green).



	$P(y\|\text{do}(x))$	l_{cm}	h_{cm}	l_{bp}	h_{bp}	l_{nb}	h_{nb}
$X = 0, Y = 0$	0.8934	**0.8157**	**0.9070**	0.5495	0.9206	0.6984	0.9478
$X = 0, Y = 1$	0.1066	**0.0930**	**0.1843**	0.0794	0.4505	0.0522	0.3016
$X = 1, Y = 0$	0.8964	**0.7391**	**0.9185**	0.2944	0.9591	0.4416	0.9970
$X = 1, Y = 1$	0.1036	**0.0815**	**0.2609**	0.0409	0.7056	0.0030	0.5584

Table 1: Causal bounds $[l, h]$ over the interventional distribution $P(y|\text{do}(x))$. The optimal bounds are marked in **bold**.