Differentiable Causal Discovery
with observational and interventional data


Philippe Brouillard *[1]


Sébastien Lachapelle *[1]


Alexandre Lacoste [2]


Simon Lacoste-Julien [1]


Alexandre Drouin [2]

[1] Mila & DIRO, Université de Montréal [2] Element AI / ServiceNow * Equal contribution

OATML, University of Oxford

February 16, 2021

ELEMENT AI
a ServiceNow Company

Mila

Université de Montréal

Outline

Introduction and Motivation

The limits of statistical association

Consider the relationships between altitude (A) and temperature (T)

Example taken from [Peters et al., 2017]

$$P(A, T) = P(A|T)P(T)$$
$$= P(T|A)P(A)$$

If altitude $\uparrow$ then temperature $\downarrow$

The limits of statistical association

Consider the relationships between altitude (A) and temperature (T)

Example taken from [Peters et al., 2017]

$$P(A, T) = P(A|T)P(T)$$
$$= P(T|A)P(A)$$

If altitude $\downarrow$ then temperature $\uparrow$

The limits of statistical association

Consider the relationships between altitude (A) and temperature (T)

Example taken from [Peters et al., 2017]

$$P(A, T) = P(A|T)P(T)$$
$$= P(T|A)P(A)$$

Will cooling house #1 make it climb the mountain?

The limits of statistical association

Consider the relationships between altitude (A) and temperature (T)

Example taken from [Peters et al., 2017]

$$P(A, T) = P(A|T)P(T)$$
$$= P(T|A)P(A)$$

Will pushing house #2 down the mountain change its temperature?

Why care about causal relationships: Simpson's paradox

Recovery of kidney stone patients

	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

Example taken from Julious & Mullee [1994]

Why care about causal relationships: Simpson's paradox

Recovery of kidney stone patients

	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

Example taken from Julious & Mullee [1994]

- Small stones: treatment **a** more effective
- Large stones: treatment **a** more effective

Why care about causal relationships: Simpson's paradox

Recovery of kidney stone patients

	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

Example taken from Julious & Mullee [1994]

- Small stones: treatment **a** more effective
- Large stones: treatment **a** more effective

Why care about causal relationships: Simpson's paradox

Recovery of kidney stone patients

	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

Example taken from Julious & Mullee [1994]

- Small stones: treatment **a** more effective
- Large stones: treatment **a** more effective

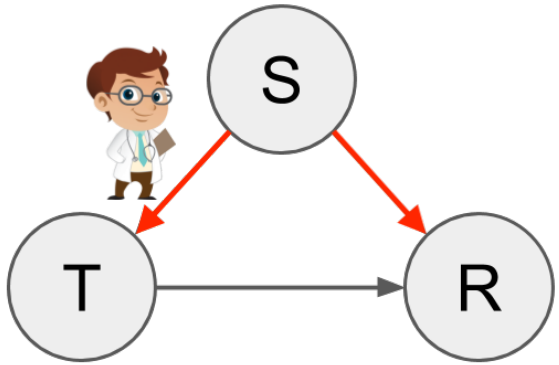
All patients: treatment **b** more effective!

Simpson's paradox: what's really going on?

T = Treatment $\in \{A, B\}$

S = Stone size $\in \{\text{small}, \text{large}\}$

R = Patient recovered $\in \{0, 1\}$



	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

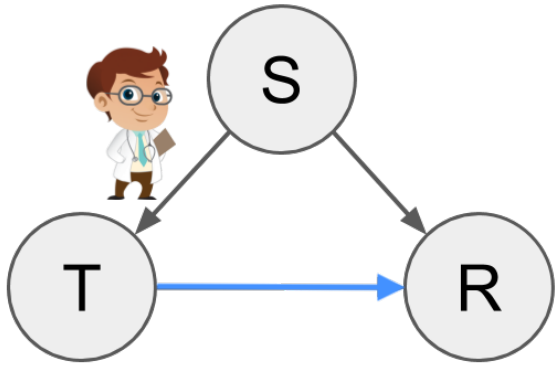
Confounding: patients with small stones are more likely to receive treatment b and also more likely to have a good outcome.

Simpson's paradox: what's really going on?

T = Treatment $\in \{A, B\}$

S = Stone size $\in \{\text{small}, \text{large}\}$

R = Patient recovered $\in \{0, 1\}$

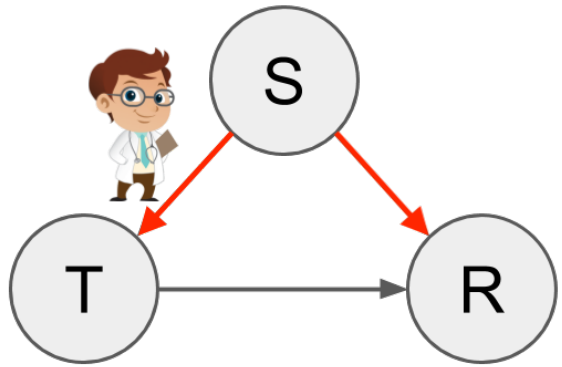


	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

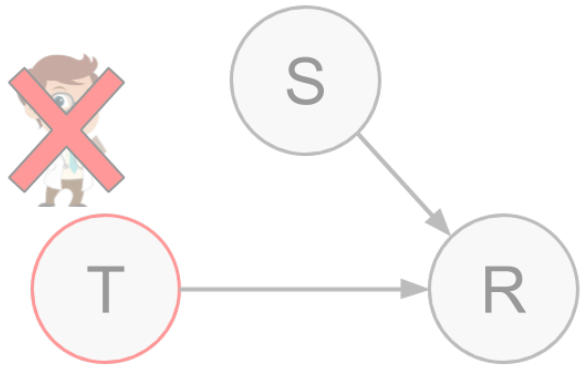
Confounding: patients with small stones are more likely to receive treatment b and also more likely to have a good outcome.

Causal vs non-causal questions

- Non-causal: What is the probability of recovery given that the doctor assigned treatment A?
  $\hookrightarrow P(R = 1 \mid T = A)$

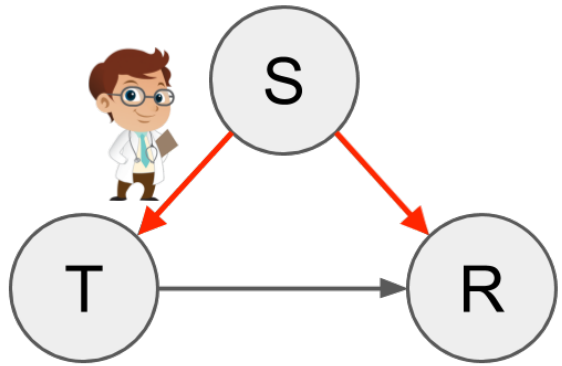


- Causal: What's the probability of recovery if treatment A is used irrespective of the size of the stones?
  $\hookrightarrow P(R = 1 \mid do(T = A))$

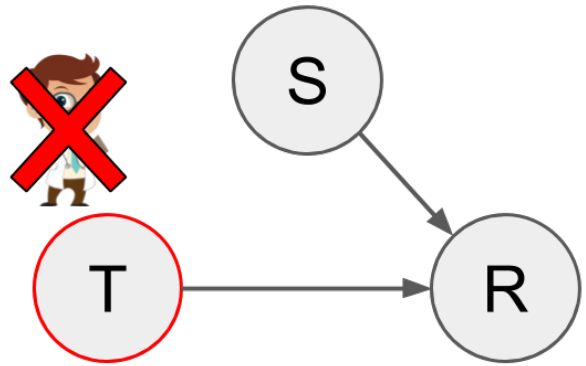


Causal vs non-causal questions

- Non-causal: What is the probability of recovery given that the doctor assigned treatment A?
  $\hookrightarrow$ $P(R = 1 \mid T = A)$



- Causal: What's the probability of recovery if treatment A is used irrespective of the size of the stones?
  $\hookrightarrow$ $P(R = 1 \mid do(T = A))$

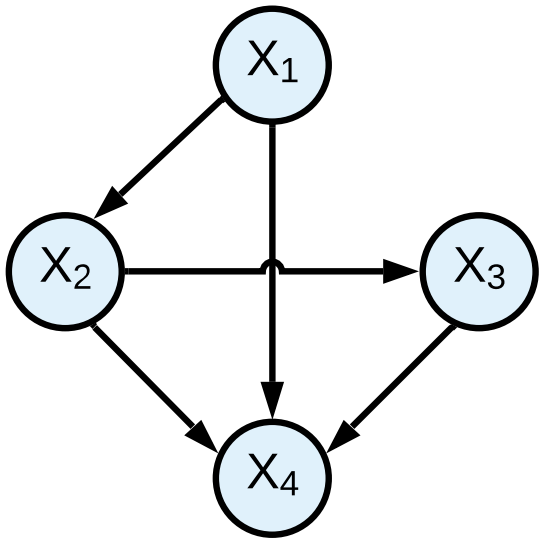


Causal Bayesian networks

- Random vector $X = (X_1, ..., X_d)$
- Let $\mathcal{G}$ be a directed acyclic graph (DAG)
  - d vertices (one per X_i)
  - edges indicate causal relationships
- Encodes (conditional) independence constraints (via *d-separation*, see Koller & Friedman [2009])
- Distribution P_X: $p(X) = \prod_{i=1}^{d} p(X_i | X_{\pi_i^{\mathcal{G}}})$, where $\pi_i^{\mathcal{G}}$ = parents of i in $\mathcal{G}$

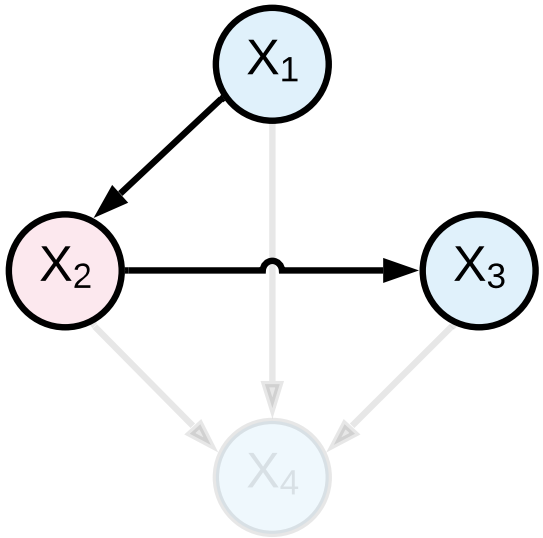
Causal Bayesian networks

- Random vector $X = (X_1, ..., X_d)$
- Let $\mathcal{G}$ be a directed acyclic graph (DAG)
  - d vertices (one per X_i)
  - edges indicate causal relationships
- Encodes (conditional) independence constraints (via *d-separation*, see Koller & Friedman [2009])
- Distribution P_X: $p(X) = \prod_{i=1}^{d} p(X_i | X_{\pi_i^{\mathcal{G}}})$, where $\pi_i^{\mathcal{G}}$ = parents of i in $\mathcal{G}$



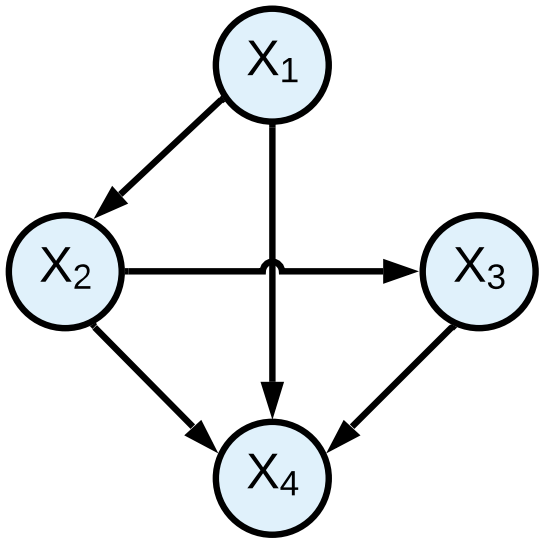
Causal Bayesian networks

- Random vector $X = (X_1, ..., X_d)$
- Let $\mathcal{G}$ be a directed acyclic graph (DAG)
  - d vertices (one per X_i)
  - edges indicate causal relationships
- Encodes (conditional) independence constraints (via *d-separation*, see Koller & Friedman [2009])
- Distribution P_X: $p(X) = \prod_{i=1}^{d} p(X_i | X_{\pi_i^{\mathcal{G}}})$, where $\pi_i^{\mathcal{G}}$ = parents of i in $\mathcal{G}$



Causal Bayesian networks

- Random vector $X = (X_1, ..., X_d)$
- Let $\mathcal{G}$ be a directed acyclic graph (DAG)
  - d vertices (one per X_i)
  - edges indicate causal relationships
- Encodes (conditional) independence constraints (via *d-separation*, see Koller & Friedman [2009])
- Distribution P_X: $p(X) = \prod_{i=1}^{d} p(X_i | X_{\pi_i^{\mathcal{G}}})$, where $\pi_i^{\mathcal{G}}$ = parents of i in $\mathcal{G}$

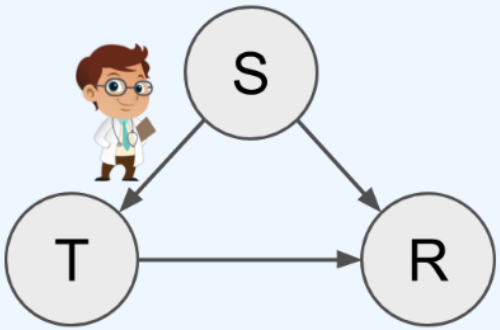


Interventions

Intervening on the treatment T

$T =$ Treatment $\in \{A, B\}$
$S =$ Stone size $\in \{\text{small}, \text{large}\}$
$R =$ Patient recovered $\in \{0, 1\}$

Observations



$$p(S)p(T \mid S)p(R \mid S, T)$$

Observational data: may contain biases

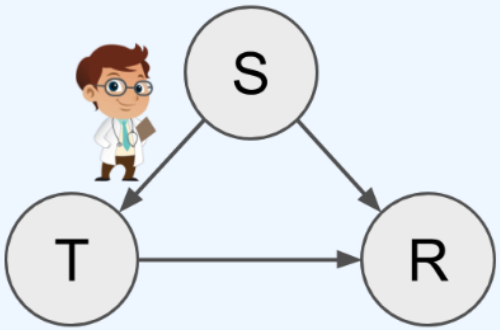
$$P(R = 1 \mid T = A) \neq P(R = 1 \mid do(T = A))$$

Interventions

Intervening on the treatment T

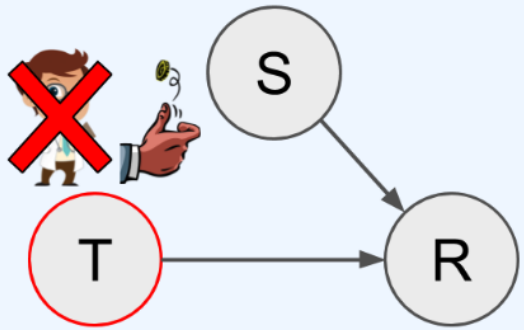
$T =$ Treatment $\in \{A, B\}$
$S =$ Stone size $\in \{\text{small}, \text{large}\}$
$R =$ Patient recovered $\in \{0, 1\}$

Observations



$p(S)p(T \mid S)p(R \mid S, T)$

Perfect intervention



$p(S)\tilde{p}(T)p(R \mid S, T)$

Perfect intervention: edges into T are removed (e.g., via randomization)

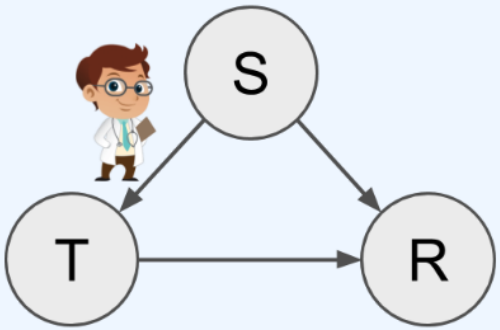
$$P(R = 1 \mid T = A) = P(R = 1 \mid do(T = A))$$

Interventions

Intervening on the treatment T

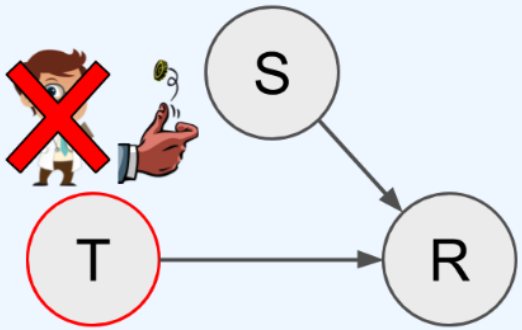
$T =$ Treatment $\in \{A, B\}$
$S =$ Stone size $\in \{\text{small}, \text{large}\}$
$R =$ Patient recovered $\in \{0, 1\}$

Observations



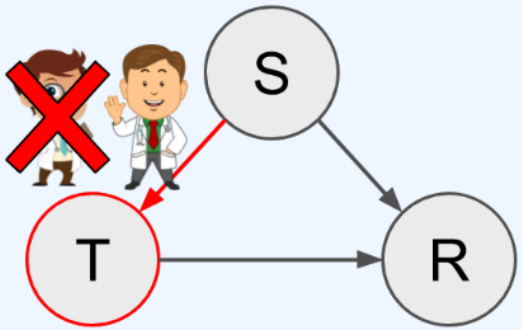
$$p(S)p(T \mid S)p(R \mid S, T)$$

Perfect intervention



$$p(S)\tilde{p}(T)p(R \mid S, T)$$

Imperfect intervention



$$p(S)\tilde{p}(T \mid S)p(R \mid S, T)$$

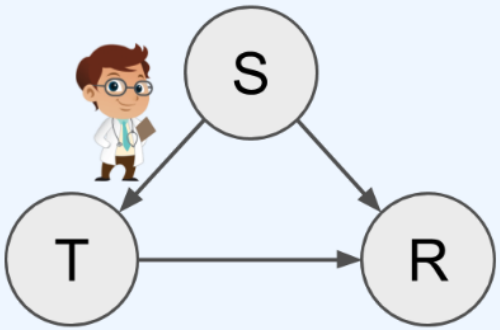
Imperfect intervention: incoming edges are preserved, conditionals are changed.

Interventions

Intervening on the treatment T

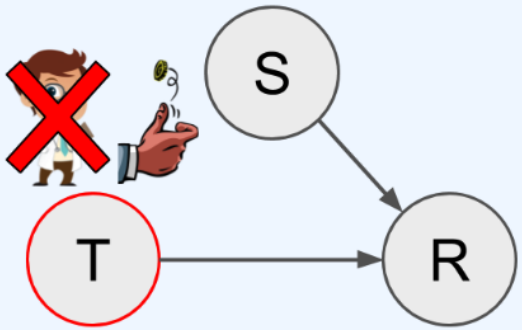
$T =$ Treatment $\in \{A, B\}$
$S =$ Stone size $\in \{\text{small}, \text{large}\}$
$R =$ Patient recovered $\in \{0, 1\}$

Observations



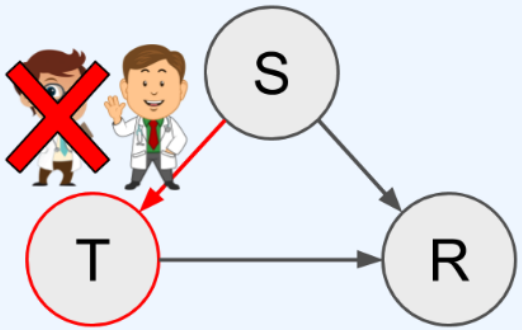
$$p(S)p(T \mid S)p(R \mid S, T)$$

Perfect intervention



$$p(S)\tilde{p}(T)p(R \mid S, T)$$

Imperfect intervention



$$p(S)\tilde{p}(T \mid S)p(R \mid S, T)$$

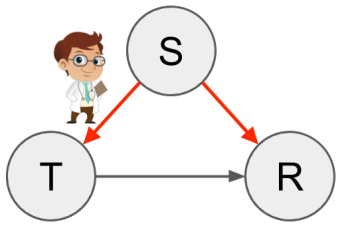
⚠ **Important:** notice how conditionals that are not under intervention are invariant across distributions (a.k.a, modularity/autonomy).

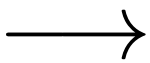
Causal inference from observational data

- **Objective:** estimate an causal quantity $P(R = 1 \mid do(T = A))$

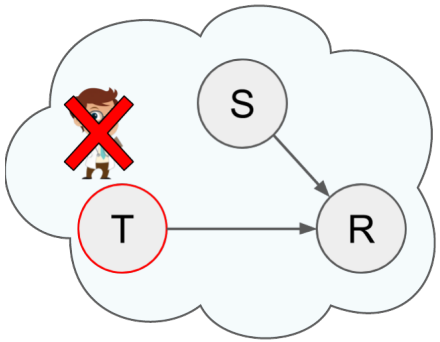
  ↳ randomization may not be possible (e.g., unethical)

Observational distribution





Interventional distribution



- **Identification:** transforming a causal quantity into a purely statistical one
  - do-calculus [Pearl, 2009] automates this and is even complete [Huang & Valtorta, 2006; Shpitser & Pearl, 2006]

- **Example:** *covariate adjustment*

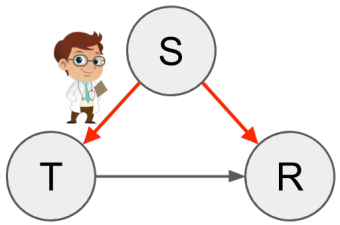
$$P(R = 1 \mid do(T = A)) = \sum_{s \in \{\text{small,large}\}} P(R = 1 \mid T = A, S = s)\, P(S = s)$$

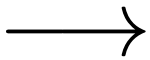
Causal inference from observational data

- **Objective:** estimate an causal quantity $P(R = 1 \mid do(T = A))$

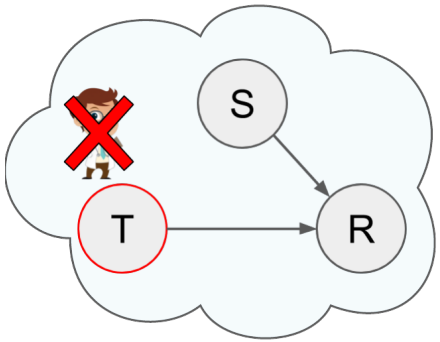
  ↳ randomization may not be possible (e.g., unethical)

Observational distribution





Interventional distribution



- **Identification:** transforming a causal quantity into a purely statistical one
  - do-calculus [Pearl, 2009] automates this and is even complete [Huang & Valtorta, 2006; Shpitser & Pearl, 2006]

- **Example:** *covariate adjustment*

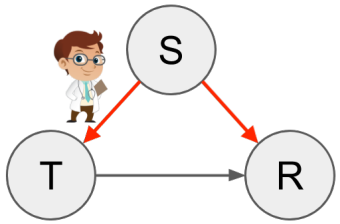
$$P(R = 1 \mid do(T = A)) = \sum_{s \in \{\text{small,large}\}} P(R = 1 \mid T = A, S = s)\, P(S = s)$$

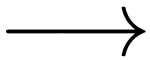
Causal inference from observational data

- **Objective:** estimate an causal quantity $P(R = 1 \mid do(T = A))$

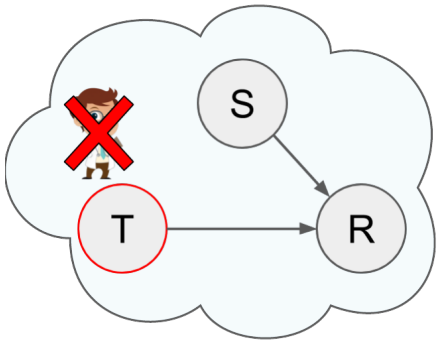
  ↳ randomization may not be possible (e.g., unethical)

Observational distribution





Interventional distribution



- **Identification:** transforming a causal quantity into a purely statistical one
  - do-calculus [Pearl, 2009] automates this and is even complete [Huang & Valtorta, 2006; Shpitser & Pearl, 2006]

- **Example:** *covariate adjustment*

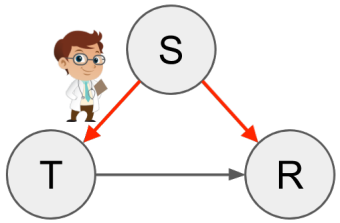
$$P(R = 1 \mid do(T = A)) = \sum_{s \in \{\text{small,large}\}} P(R = 1 \mid T = A, S = s)\ P(S = s)$$

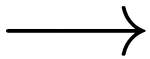
Causal inference from observational data

- **Objective:** estimate an causal quantity $P(R = 1 \mid do(T = A))$

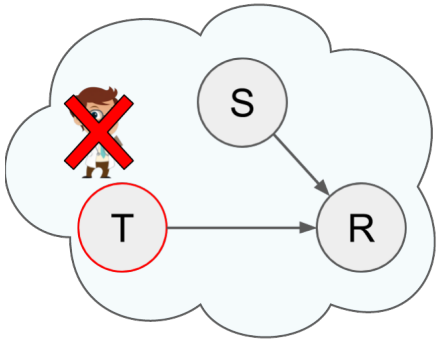
  ↳ randomization may not be possible (e.g., unethical)

Observational distribution





Interventional distribution

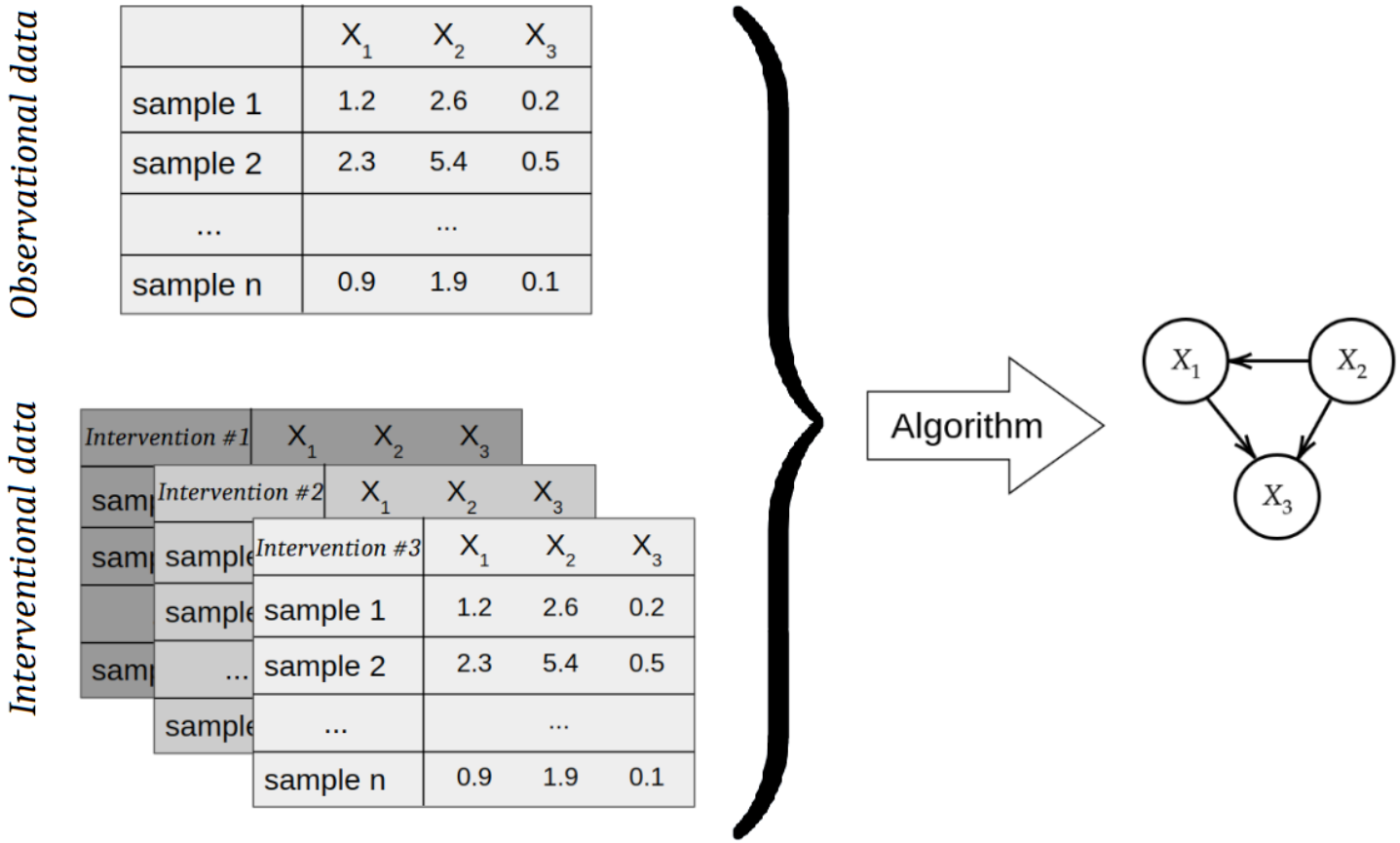


- **Identification:** transforming a causal quantity into a purely statistical one
  - do-calculus [Pearl, 2009] automates this and is even complete [Huang & Valtorta, 2006; Shpitser & Pearl, 2006]

What if you don't know the causal graph?

Causal Discovery

Problem statement



Common assumptions

To make this possible, we need to make assumptions

- **Causal sufficiency:** no hidden confounding variables
- **Markov property:** *d-separation* in the graph implies *conditional independence*

  $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{P_X} X_2 \,|\, Z$
- **Faithfulness:** *conditional independence* implies *d-separation* in the graph

  $X_1 \perp\!\!\!\perp_{P_X} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z$

Common assumptions

To make this possible, we need to make assumptions

- **Causal sufficiency:** no hidden confounding variables
- **Markov property:** *d-separation* in the graph implies *conditional independence*

  $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{P_X} X_2 \,|\, Z$
- **Faithfulness:** *conditional independence* implies *d-separation* in the graph

  $X_1 \perp\!\!\!\perp_{P_X} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z$

Common assumptions

To make this possible, we need to make assumptions

- **Causal sufficiency:** no hidden confounding variables
- **Markov property:** *d-separation* in the graph implies *conditional independence*

  $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{P_X} X_2 \,|\, Z$
- **Faithfulness:** *conditional independence* implies *d-separation* in the graph

  $X_1 \perp\!\!\!\perp_{P_X} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z$

Common assumptions

To make this possible, we need to make assumptions

- **Causal sufficiency:** no hidden confounding variables
- **Markov property:** *d-separation* in the graph implies *conditional independence*

  $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{P_X} X_2 \,|\, Z$
- **Faithfulness:** *conditional independence* implies *d-separation* in the graph

  $X_1 \perp\!\!\!\perp_{P_X} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z$

Common assumptions

To make this possible, we need to make assumptions

- **Causal sufficiency:** no hidden confounding variables
- **Markov property:** *d-separation* in the graph implies *conditional independence*

  $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{P_X} X_2 \,|\, Z$
- **Faithfulness:** *conditional independence* implies *d-separation* in the graph

  $X_1 \perp\!\!\!\perp_{P_X} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z$

These last two assumptions guarantee an equivalence between properties of the data and properties of the graph

Score-based causal discovery

- **Idea:** find the DAG that maximizes a score function ($\mathcal{S}$)
  - E.g., data likelihood + sparsity prior
  - Consistency: need to demonstrate that the score leads to the true solution

$$\hat{\mathcal{G}} \in \underset{\mathcal{G} \in DAG}{\arg\max}\ \mathcal{S}(\mathcal{G})$$

- **Problem:** search space grows superexponentially with variables

p	number of DAGs with p nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	9942119532215951589522891459235452451655502687858830501478 3
19	3327719012271075917361775733112611258835830762584219025835467735 05
20	2344880451051088988152559855229099188899081192234291298795803236068491263

- **Examples:** Greedy Equivalence Search [Chickering, 2003], DAG with NO TEARS [Zheng et al., 2018]

Score-based causal discovery

- **Idea:** find the DAG that maximizes a score function ($\mathcal{S}$)
  - E.g., data likelihood + sparsity prior
  - Consistency: need to demonstrate that the score leads to the true solution

$$\hat{\mathcal{G}} \in \underset{\mathcal{G} \in DAG}{\arg\max}\, \mathcal{S}(\mathcal{G})$$

- **Problem:** search space grows superexponentially with variables

p	number of DAGs with p nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	9942119532215951589522891459235452451655502687858830501478 3
19	3327719012271075917361775733112611258835830762584219025835467735 05
20	2344880451051088988152559855229099188899081192234291298795803236068491263

- **Examples:** Greedy Equivalence Search [Chickering, 2003], DAG with NO TEARS [Zheng et al., 2018]

Score-based causal discovery

- **Idea:** find the DAG that maximizes a score function ($\mathcal{S}$)
  - E.g., data likelihood + sparsity prior
  - Consistency: need to demonstrate that the score leads to the true solution

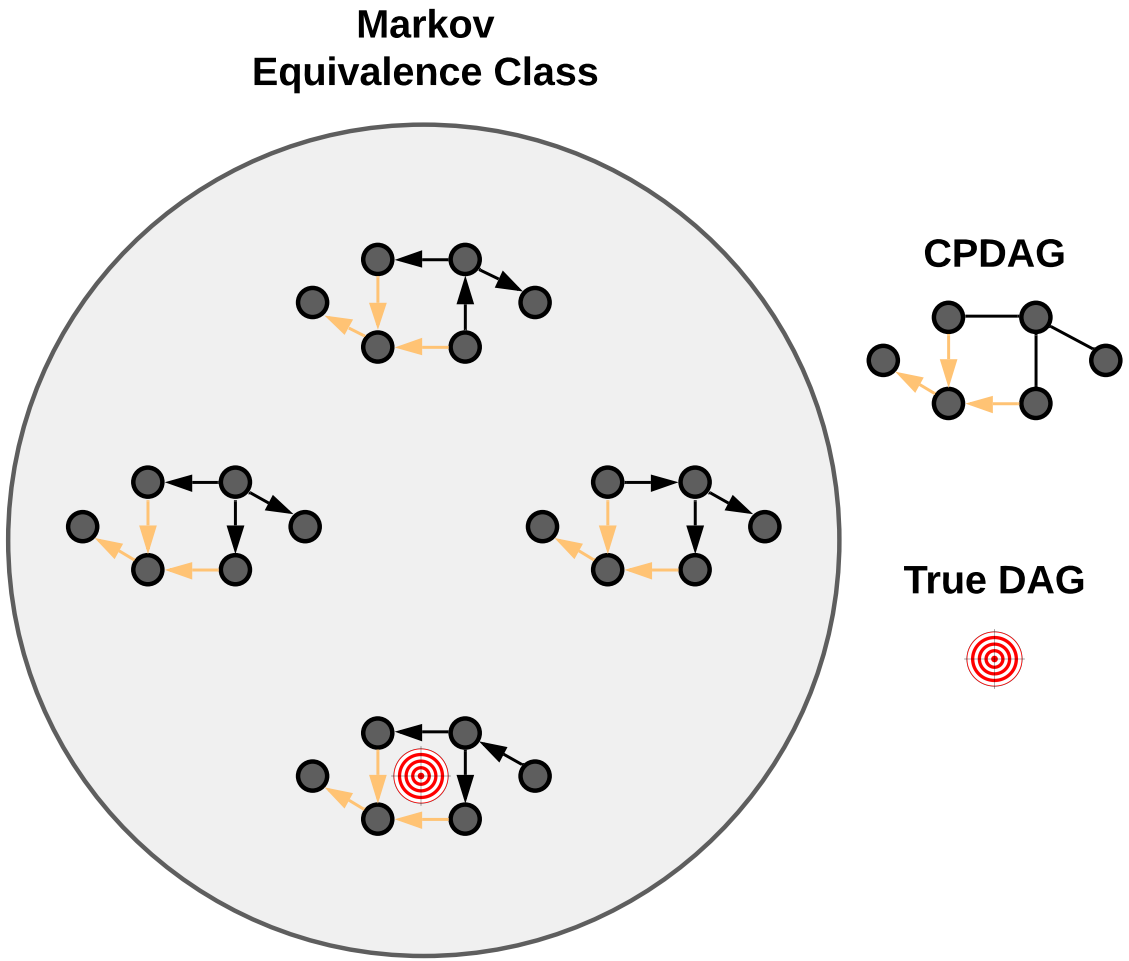
$$\hat{\mathcal{G}} \in \underset{\mathcal{G} \in DAG}{\arg\max}\, \mathcal{S}(\mathcal{G})$$

- **Problem:** search space grows superexponentially with variables

p	number of DAGs with p nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	316034593964189176074 25
12	52193965134382940502050 4063
13	18676600744432035186664816926721
14	14394281410443983349417907198395351 03
15	23772526555341035499218021828637671925 3505
16	83756670773733202876993030479964122352231 38303
17	62707921196923889899446452602494921906963551482675201
18	99421195322159515895228914592354524516555026878588305014783
19	33277190122710759173617757331126112588358307625842190258354677 3505
20	23448804510510889881525598552290991888990811922342912987958032360684912 63

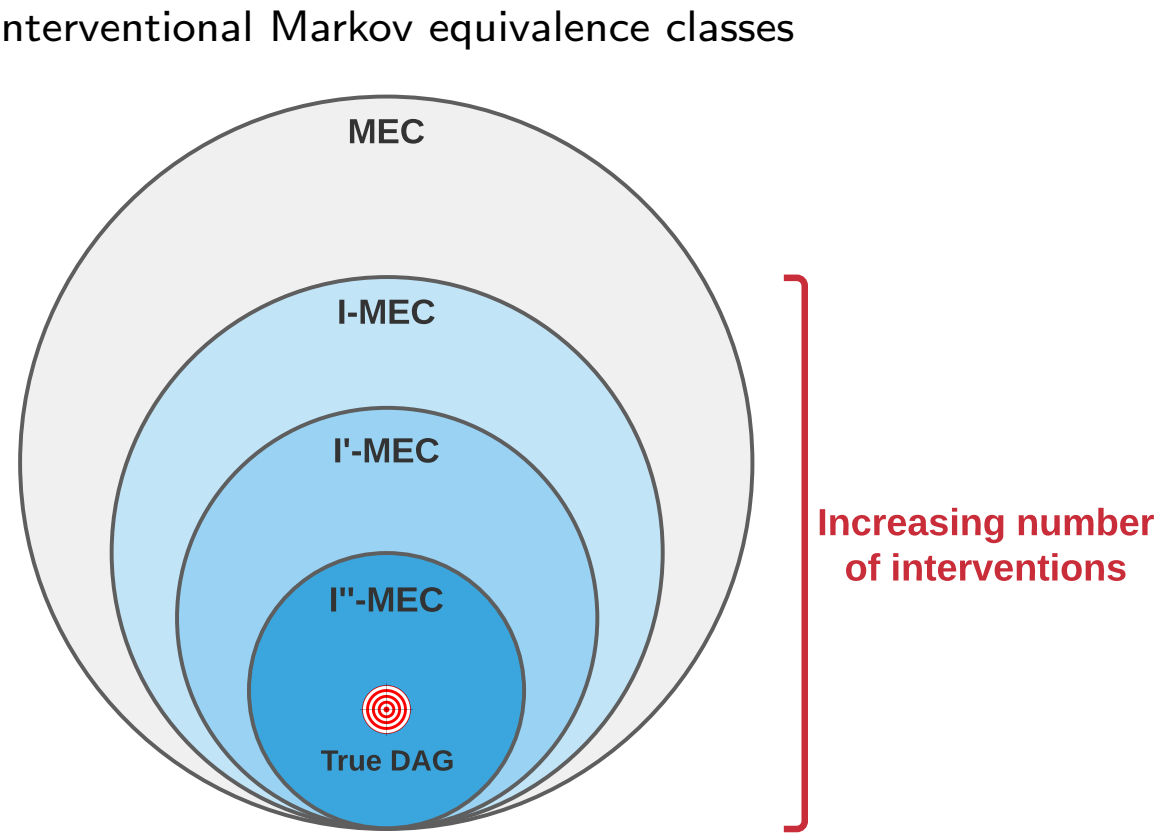
- **Examples:** Greedy Equivalence Search [Chickering, 2003], DAG with NO TEARS [Zheng et al., 2018]

Identifiability of the DAG



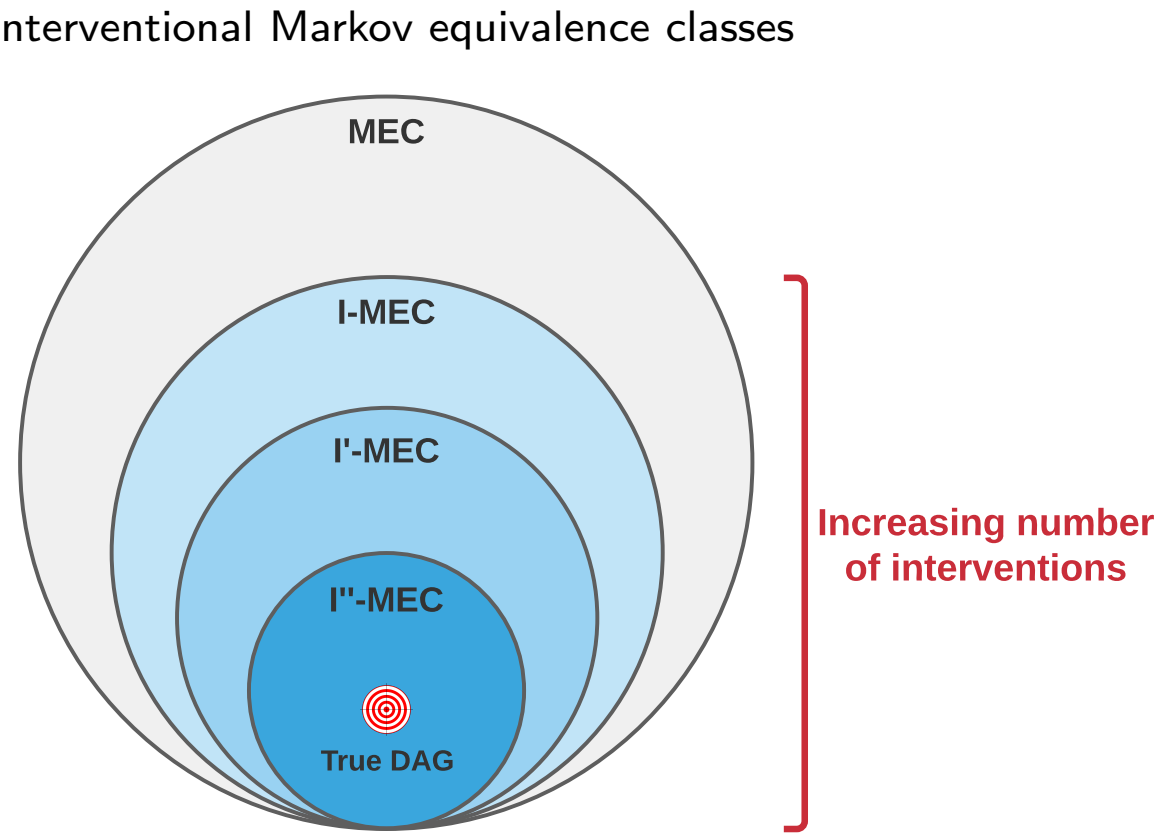
Without making more assumptions, observational data only allows identification up to a Markov equivalence class (MEC) [Verma & Pearl, 1991]

Can you shrink the equivalence class?



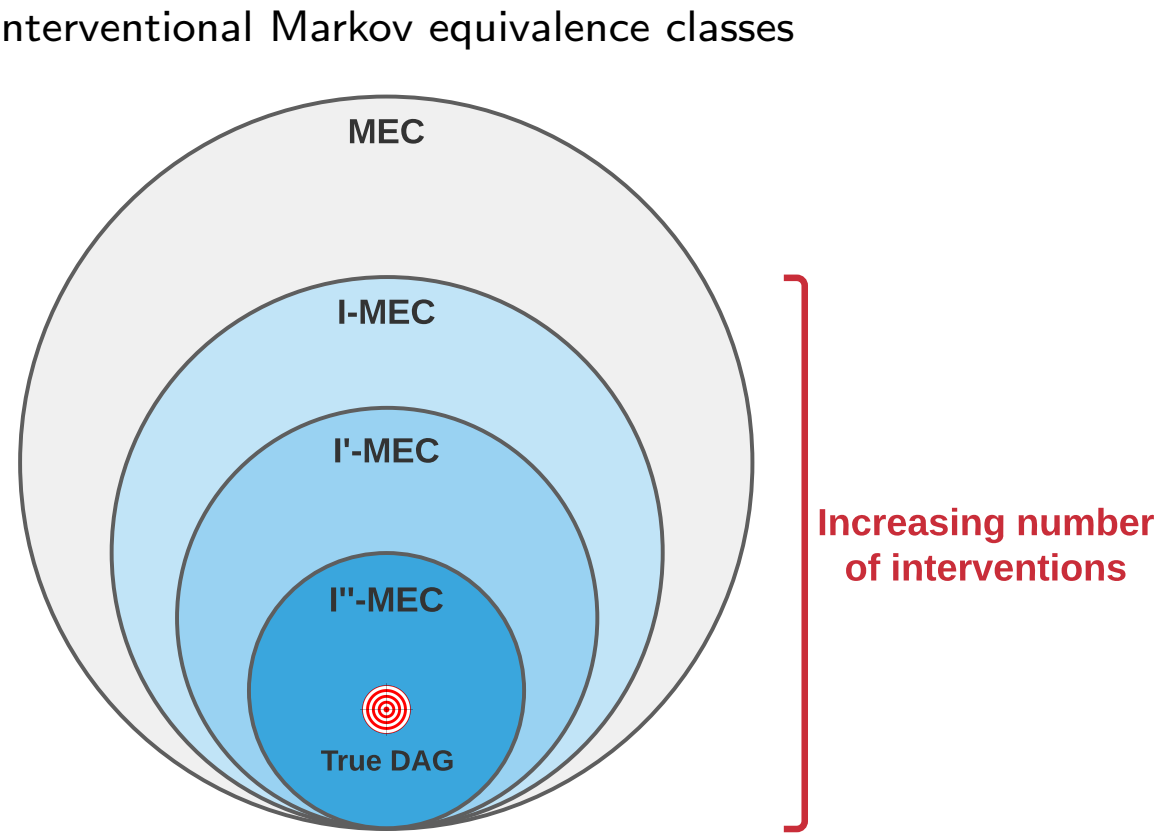
- An I-MEC is a subset of the MEC (see Eberhardt et al. [2005])
- **Example:** gene knockout/knockdown experiments in biology [Dixit et al., 2016]

Can you shrink the equivalence class?



- An I-MEC is a subset of the MEC (see Eberhardt et al. [2005])
- **Example:** gene knockout/knockdown experiments in biology [Dixit et al., 2016]

Can you shrink the equivalence class?



- An I-MEC is a subset of the MEC (see Eberhardt et al. [2005])
- **Example:** gene knockout/knockdown experiments in biology [Dixit et al., 2016]

Differentiable Causal Discovery with Interventional Data

Differentiable Causal Discovery from Interventional Data

Philippe Brouillard* Mila, Université de Montréal

Sébastien Lachapelle* Mila, Université de Montréal

Alexandre Lacoste Element AI

Simon Lacoste-Julien Mila, Université de Montréal Canada CIFAR AI Chair

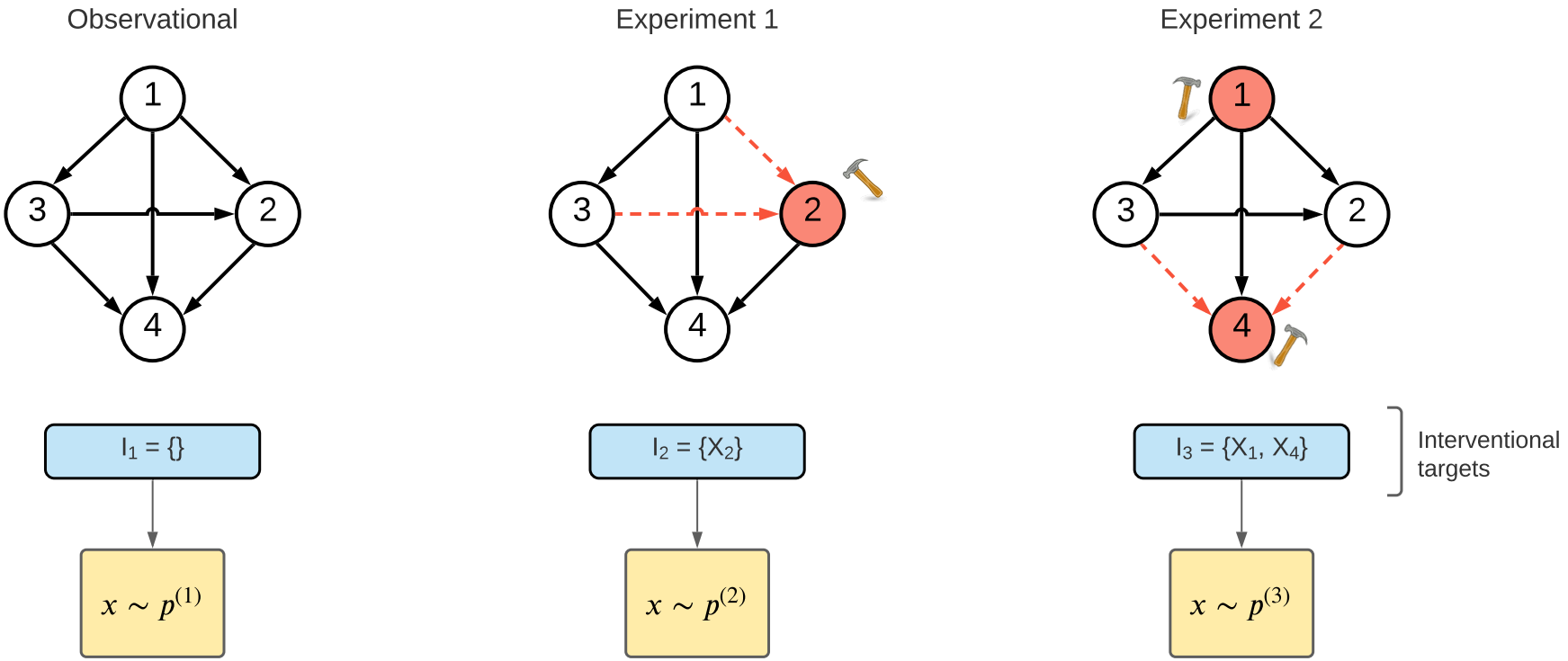
Alexandre Drouin Element AI

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

DCDI Fact Sheet:

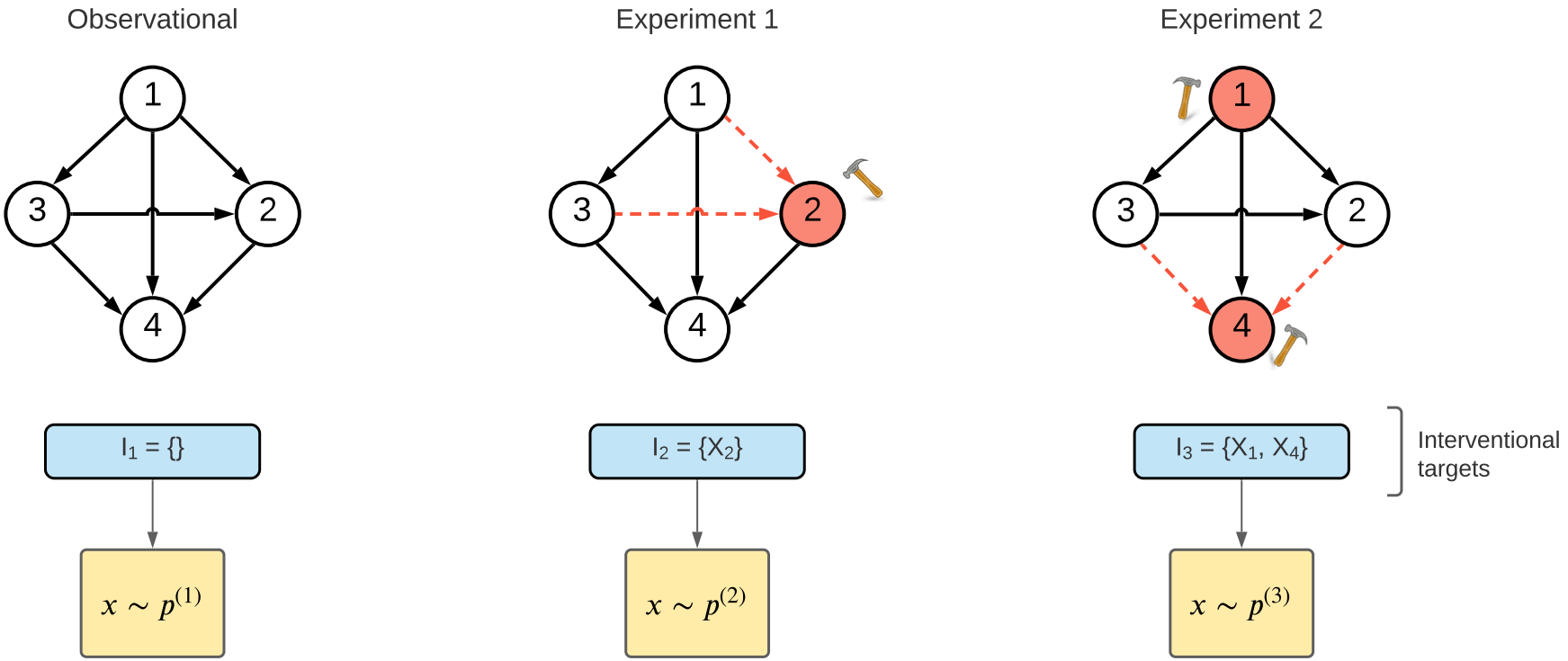
- **Type:** score-based
- **Search strategy:** continuous-constrained optimization [Zheng et al., 2018]
- **Data:** observational and interventional (perfect/imperfect)
- **Theoretical guarantees:** consistency in the limit of infinite data

Interventional distributions and the invariance property



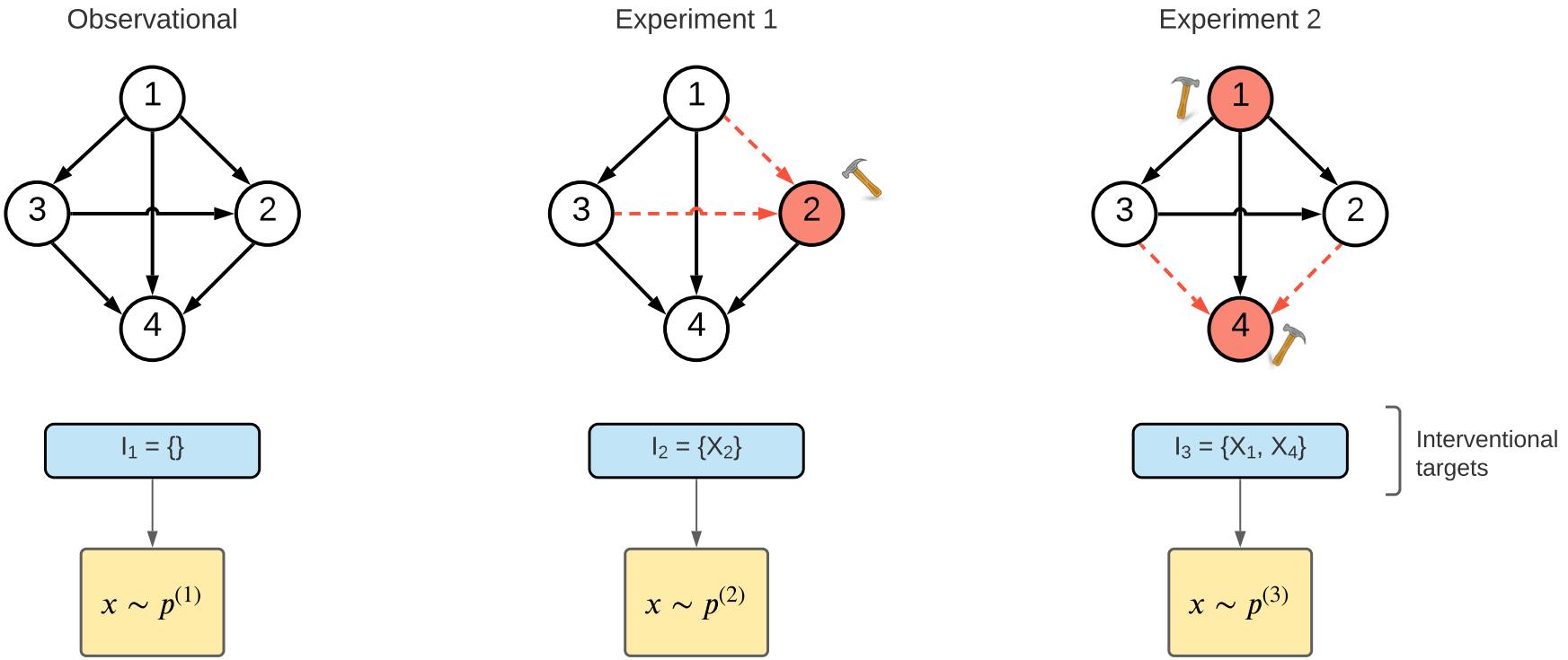
$$p^{(1)}(x_1, ..., x_4) := p^{(1)}(x_1)p^{(1)}(x_3|x_1)p^{(1)}(x_2|x_1, x_3)p^{(1)}(x_4|x_1, x_2, x_3)$$

Interventional distributions and the invariance property



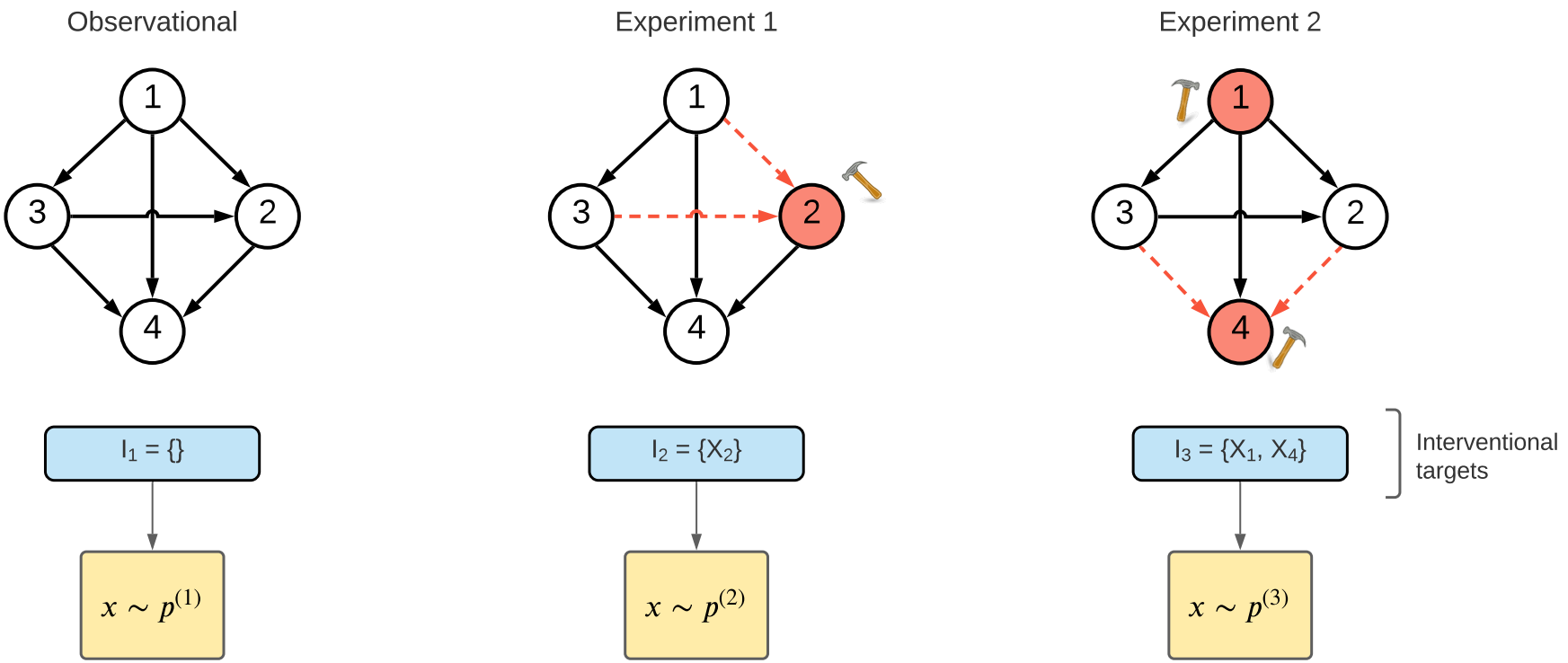
$$p^{(2)}(x_1, ..., x_4) := p^{(1)}(x_1)p^{(1)}(x_3|x_1)p^{(2)}(x_2|x_1, x_3)p^{(1)}(x_4|x_1, x_2, x_3)$$

Interventional distributions and the invariance property



$$p^{(3)}(x_1, ..., x_4) := p^{(3)}(x_1)p^{(1)}(x_3|x_1)p^{(1)}(x_2|x_1, x_3)p^{(3)}(x_4|x_1, x_2, x_3)$$

Interventional distributions and the invariance property



$$p^{(k)}(x_1, ..., x_d) := \prod_{j \notin I_k} p_j^{(1)}(x_j | x_{\pi_j^{\mathcal{G}}}) \prod_{j \in I_k} p_j^{(k)}(x_j | x_{\pi_j^{\mathcal{G}}})$$

DCDI: problem setting and notation

- We observe d variables which are *causally sufficient*, i.e. no hidden confounders.

$$\text{Causal DAG} = \underbrace{G = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}}_{\text{Adjacency matrix}} \in \{0,1\}^{d \times d}$$

- We have K, potentially **imperfect**, interventions which can target multiple variables simultaneously.

$$\underbrace{I = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}}_{\text{Intervention matrix}} \in \{0,1\}^{K \times d}$$

DCDI: problem setting and notation

- We observe d variables which are *causally sufficient*, i.e. no hidden confounders.

$$\text{Causal DAG} = \underbrace{G = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}}_{\text{Adjacency matrix}} \in \{0,1\}^{d \times d}$$

- We have K, potentially **imperfect**, interventions which can target multiple variables simultaneously.

$$\underbrace{I = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}}_{\text{Intervention matrix}} \in \{0,1\}^{K \times d}$$

DCDI: problem setting and notation

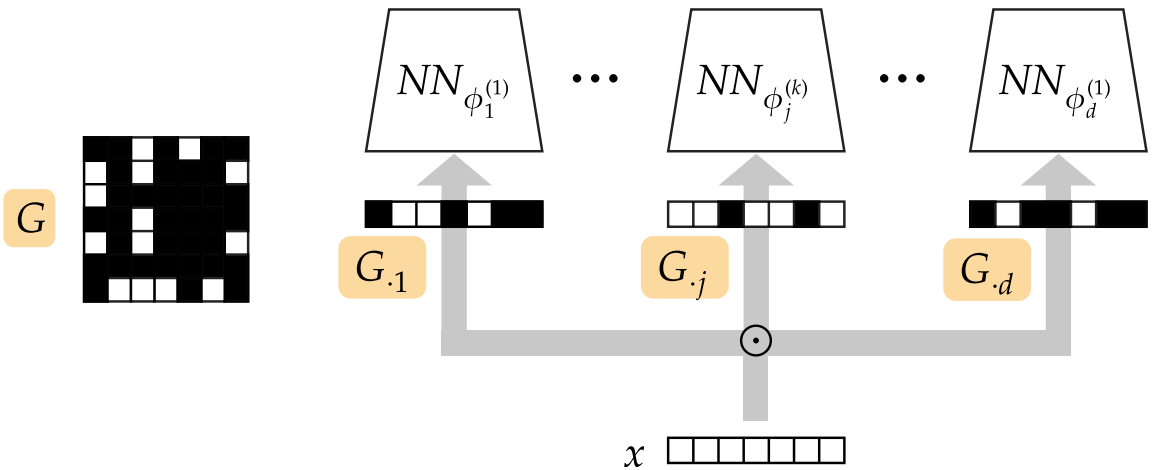
- We observe d variables which are *causally sufficient*, i.e. no hidden confounders.

$$\text{Causal DAG} = \underbrace{G = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}}_{\text{Adjacency matrix}} \in \{0,1\}^{d \times d}$$

- We have K, potentially **imperfect**, interventions which can target multiple variables simultaneously.

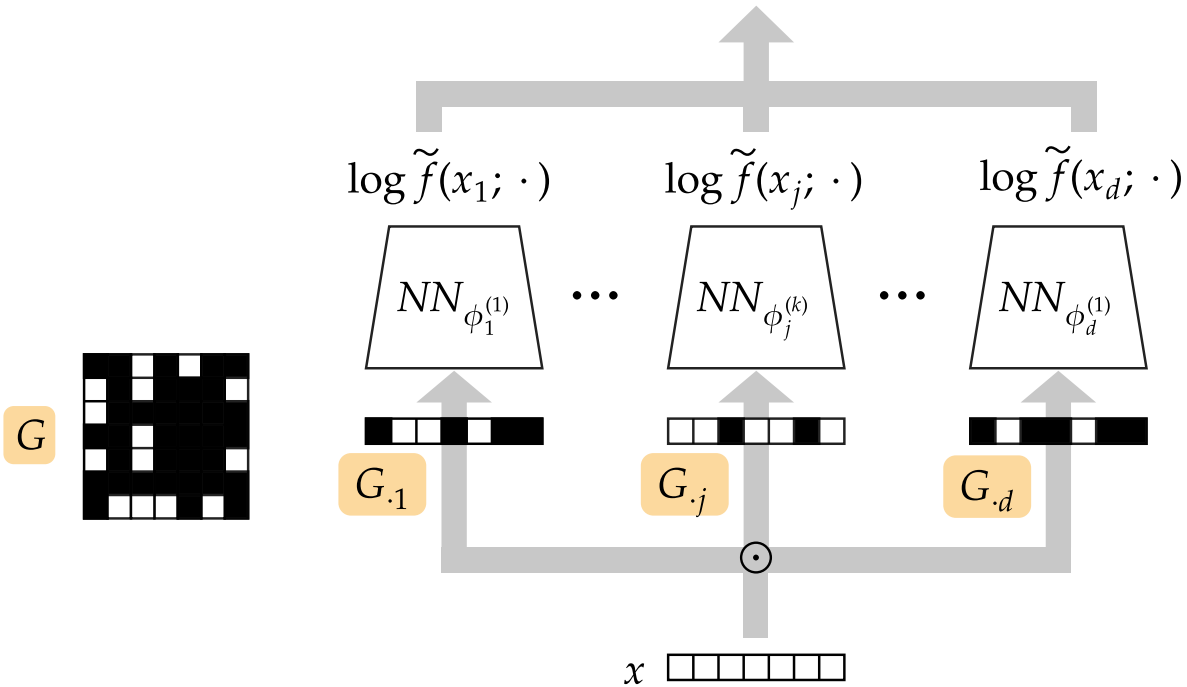
$$\underbrace{I = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}}_{\text{Intervention matrix}} \in \{0,1\}^{K \times d}$$

DCDI: model architecture



The graph adjacency matrix acts as a mask that filters the input variables.

DCDI: model architecture



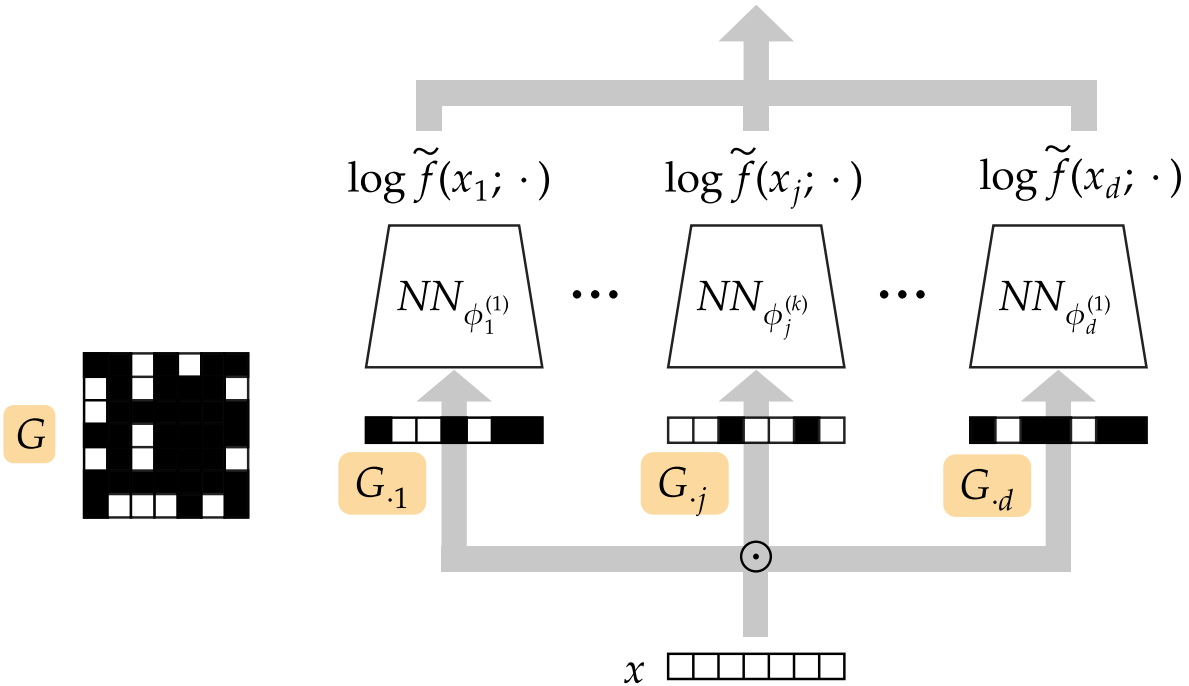
Each conditional distribution is estimated by a distinct neural network.

DCDI: model architecture

Observational parameter

Interventional parameter

$$f^{(k)}(x; G, I, \phi) := \prod_{j=1}^{d} \tilde{f}(x_j; \mathrm{NN}(G_j \odot x; \overbrace{\phi_j^{(1)}}))^{1-I_{kj}} \tilde{f}(x_j; \mathrm{NN}(G_j \odot x; \overbrace{\phi_j^{(k)}}))^{I_{kj}}$$



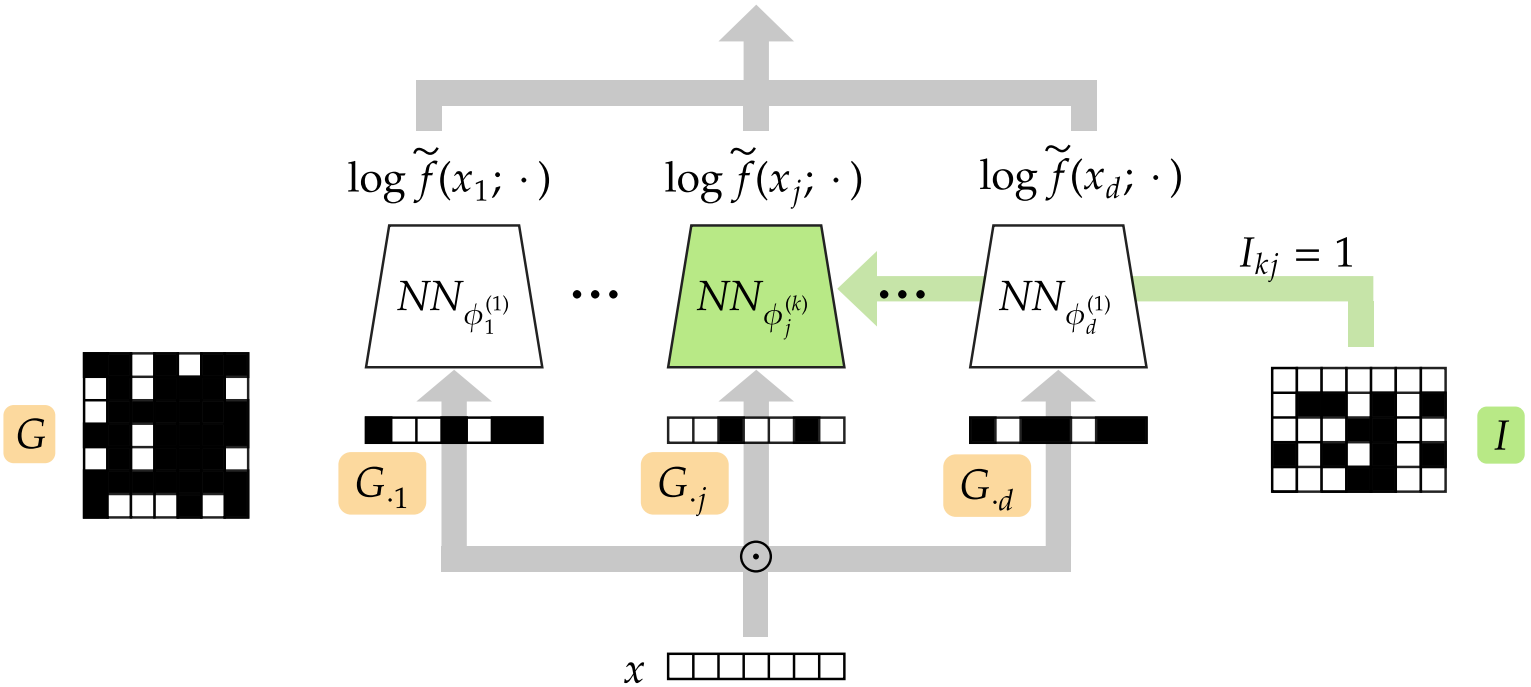
The joint likelihood is calculated as the product of conditional distributions (obs/int)

DCDI: model architecture

Observational parameter

Interventional parameter

$$f^{(k)}(x; G, I, \phi) := \prod_{j=1}^{d} \tilde{f}(x_j; \mathrm{NN}(G_j \odot x; \overbrace{\phi_j^{(1)}}))^{1-I_{kj}} \underline{\tilde{f}(x_j; \mathrm{NN}(G_j \odot x; \overbrace{\phi_j^{(k)}}))^{I_{kj}}}$$



The intervention matrix I activates the right set of parameters.

DCDI: graph scoring function (discrete)

- We suggest maximizing this score over the space of DAGs:

$$\mathcal{S}_{I^*}(G) := \sup_{\phi} \sum_{k=1}^{K} \underbrace{\mathbb{E}_{X \sim p^{(k)}}}_{k\text{th ground truth intervention}} \log f^{(k)}(X; G, I^*, \phi) - \underbrace{\lambda \|G\|_0}_{\text{Sparsity regularization}}$$

- **Search:** discrete search over DAGs $\rightarrow$ continuous-constrained opt. problem [Zheng et al., 2018]

DCDI: graph scoring function (discrete)

- We suggest maximizing this score over the space of DAGs:

$$\mathcal{S}_{I^*}(G) := \sup_{\phi} \sum_{k=1}^{K} \underbrace{\mathbb{E}_{X \sim p^{(k)}}}_{k\text{th ground truth intervention}} \log f^{(k)}(X; G, I^*, \phi) - \underbrace{\lambda \|G\|_0}_{\text{Sparsity regularization}}$$

- **Search:** discrete search over DAGs $\rightarrow$ continuous-constrained opt. problem [Zheng et al., 2018]

DCDI: theoretical justification

- $\mathcal{G}^*$ = ground-truth DAG
- I^* = ground-truth intervention matrix

$\hat{\mathcal{G}} \in \arg\max_{\mathcal{G} \in \text{DAG}} \mathcal{S}_{I^*}(\mathcal{G})$ is the estimator.

Theorem (Identification via score maximization)

Suppose $I^*_{1,:} = \emptyset$. Given that

1. *Each variable is individually targeted by an intervention;*
2. *The model has enough capacity to express the ground truth;*
3. *The regularization coefficient $\lambda > 0$ is small enough;*
4. *And some more technical assumptions, e.g. I^*-faithfulness... (See paper)*

then

$$\hat{\mathcal{G}} = \mathcal{G}^*$$

More general result

Without the first assumption, we can identify the I^*-Markov equivalence class[a] of $\mathcal{G}^*$.

[a] We use the notion of I^*-Markov equivalence of Yang et al. [2018].

DCDI: theoretical justification

- $\mathcal{G}^*$ = ground-truth DAG
- I^* = ground-truth intervention matrix

$\hat{\mathcal{G}} \in \arg\max_{\mathcal{G} \in \text{DAG}} \mathcal{S}_{I^*}(\mathcal{G})$ is the estimator.

Theorem (Identification via score maximization)

Suppose $I^*_{1,:} = \emptyset$. Given that

1. *Each variable is individually targeted by an intervention;*
2. *The model has enough capacity to express the ground truth;*
3. *The regularization coefficient $\lambda > 0$ is small enough;*
4. *And some more technical assumptions, e.g. I^*-faithfulness... (See paper)*

then

$$\hat{\mathcal{G}} = \mathcal{G}^*$$

More general result

Without the first assumption, we can identify the I^*-Markov equivalence class[a] of $\mathcal{G}^*$.

[a] We use the notion of I^*-Markov equivalence of Yang et al. [2018].

DCDI: theoretical justification

- $\mathcal{G}^*$ = ground-truth DAG
- I^* = ground-truth intervention matrix

$\hat{\mathcal{G}} \in \arg\max_{\mathcal{G} \in \text{DAG}} \mathcal{S}_{I^*}(\mathcal{G})$ is the estimator.

Theorem (Identification via score maximization)

Suppose $I^*_{1,:} = \emptyset$. Given that

1. *Each variable is individually targeted by an intervention;*
2. *The model has enough capacity to express the ground truth;*
3. *The regularization coefficient $\lambda > 0$ is small enough;*
4. *And some more technical assumptions, e.g. I^*-faithfulness... (See paper)*

then

$$\hat{\mathcal{G}} = \mathcal{G}^*$$

More general result

Without the first assumption, we can identify the I^*-Markov equivalence class[a] of $\mathcal{G}^*$.

[a] We use the notion of I^*-Markov equivalence of Yang et al. [2018].

DCDI: making the search efficient

$$\mathcal{S}_{I^*}(G) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0$$

↓ Relaxation where $G_{ij} \sim \text{Bernoulli}(\sigma(\Lambda_{ij}))$, with $\sigma(\cdot) :=$ sigmoid function

$$\hat{\mathcal{S}}_{I^*}(\Lambda) := \sup_{\phi} \mathop{\mathbb{E}}_{G \sim \sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda ||G||_0 \right]$$

↓ Optimize for Λ under acyclicity constraint

$$\sup_{\Lambda} \hat{\mathcal{S}}_{I^*}(\Lambda) \quad s.t. \quad \underbrace{\text{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

[Zheng et al., 2018]

DCDI: making the search efficient

$$\mathcal{S}_{I^*}(G) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0$$

↓

Relaxation where $G_{ij} \sim \text{Bernoulli}(\sigma(\Lambda_{ij}))$,
with $\sigma(\cdot) :=$ sigmoid function

$$\hat{\mathcal{S}}_{I^*}(\Lambda) := \sup_{\phi} \mathop{\mathbb{E}}_{G \sim \sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda ||G||_0 \right]$$

↓

Optimize for Λ under acyclicity constraint

$$\sup_{\Lambda} \hat{\mathcal{S}}_{I^*}(\Lambda) \quad s.t. \quad \underbrace{\text{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

[Zheng et al., 2018]

DCDI: making the search efficient

$$\mathcal{S}_{I^*}(G) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0$$

↓

Relaxation where $G_{ij} \sim \text{Bernoulli}(\sigma(\Lambda_{ij}))$,
with $\sigma(\cdot) :=$ sigmoid function

$$\hat{\mathcal{S}}_{I^*}(\Lambda) := \sup_{\phi} \mathop{\mathbb{E}}_{G \sim \sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0 \right]$$

↓

Optimize for Λ under acyclicity constraint

$$\sup_{\Lambda} \hat{\mathcal{S}}_{I^*}(\Lambda) \quad s.t. \quad \underbrace{\text{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

[Zheng et al., 2018]

DCDI: making the search efficient

$$\mathcal{S}_{I^*}(G) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0$$

$\downarrow$

Relaxation where $G_{ij} \sim \text{Bernoulli}(\sigma(\Lambda_{ij}))$,
with $\sigma(\cdot) :=$ sigmoid function

$$\hat{\mathcal{S}}_{I^*}(\Lambda) := \sup_{\phi} \mathop{\mathbb{E}}_{G \sim \sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0 \right]$$

$\downarrow$

Optimize for Λ under acyclicity constraint

$$\sup_{\Lambda} \hat{\mathcal{S}}_{I^*}(\Lambda) \quad s.t. \quad \underbrace{\text{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

[Zheng et al., 2018]

DCDI: making the search efficient

$$\mathcal{S}_{I^*}(G) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0$$

↓ Relaxation where $G_{ij} \sim \text{Bernoulli}(\sigma(\Lambda_{ij}))$, with $\sigma(\cdot) :=$ sigmoid function

$$\hat{\mathcal{S}}_{I^*}(\Lambda) := \sup_{\phi} \mathop{\mathbb{E}}_{G \sim \sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0 \right]$$

↓ Optimize for Λ under acyclicity constraint

$$\sup_{\Lambda} \hat{\mathcal{S}}_{I^*}(\Lambda) \quad s.t. \quad \underbrace{\text{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

[Zheng et al., 2018]

DCDI: optimization & gradient estimation

- Optimize jointly Λ and ϕ (NN parameters)

$$\max_{\phi,\Lambda} \mathop{\mathbb{E}}_{G\sim\sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X\sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda ||G||_0 \right] \text{ s.t. } \underbrace{\text{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

- **Optimization:** Augmented Lagrangian Method + RMSprop (as in Lachapelle et al. [2020])
- **Discrete sampling:** gradient w.r.t. Λ estimated via Gumbel-Softmax Straight-Through estimator [Jang et al., 2017; Maddison et al., 2017].
- Masks and/or the Gumbel-Softmax estimator were used before in causal discovery e.g., Kalainathan et al. [2018]; Ng et al. [2019]; Bengio et al. [2019]; Ke et al. [2019]

DCDI: optimization & gradient estimation

- Optimize jointly Λ and ϕ (NN parameters)

$$\max_{\phi,\Lambda} \mathop{\mathbb{E}}_{G\sim\sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X\sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda ||G||_0 \right] \text{ s.t. } \underbrace{\text{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

- **Optimization:** Augmented Lagrangian Method + RMSprop (as in Lachapelle et al. [2020])
- **Discrete sampling:** gradient w.r.t. Λ estimated via Gumbel-Softmax Straight-Through estimator [Jang et al., 2017; Maddison et al., 2017].
- Masks and/or the Gumbel-Softmax estimator were used before in causal discovery e.g., Kalainathan et al. [2018]; Ng et al. [2019]; Bengio et al. [2019]; Ke et al. [2019]

DCDI: optimization & gradient estimation

- Optimize jointly Λ and ϕ (NN parameters)

$$\max_{\phi,\Lambda} \mathop{\mathbb{E}}_{G\sim\sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X\sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda ||G||_0 \right] \text{ s.t. } \underbrace{\operatorname{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

- **Optimization:** Augmented Lagrangian Method + RMSprop (as in Lachapelle et al. [2020])
- **Discrete sampling:** gradient w.r.t. Λ estimated via Gumbel-Softmax Straight-Through estimator [Jang et al., 2017; Maddison et al., 2017].
- Masks and/or the Gumbel-Softmax estimator were used before in causal discovery e.g., Kalainathan et al. [2018]; Ng et al. [2019]; Bengio et al. [2019]; Ke et al. [2019]

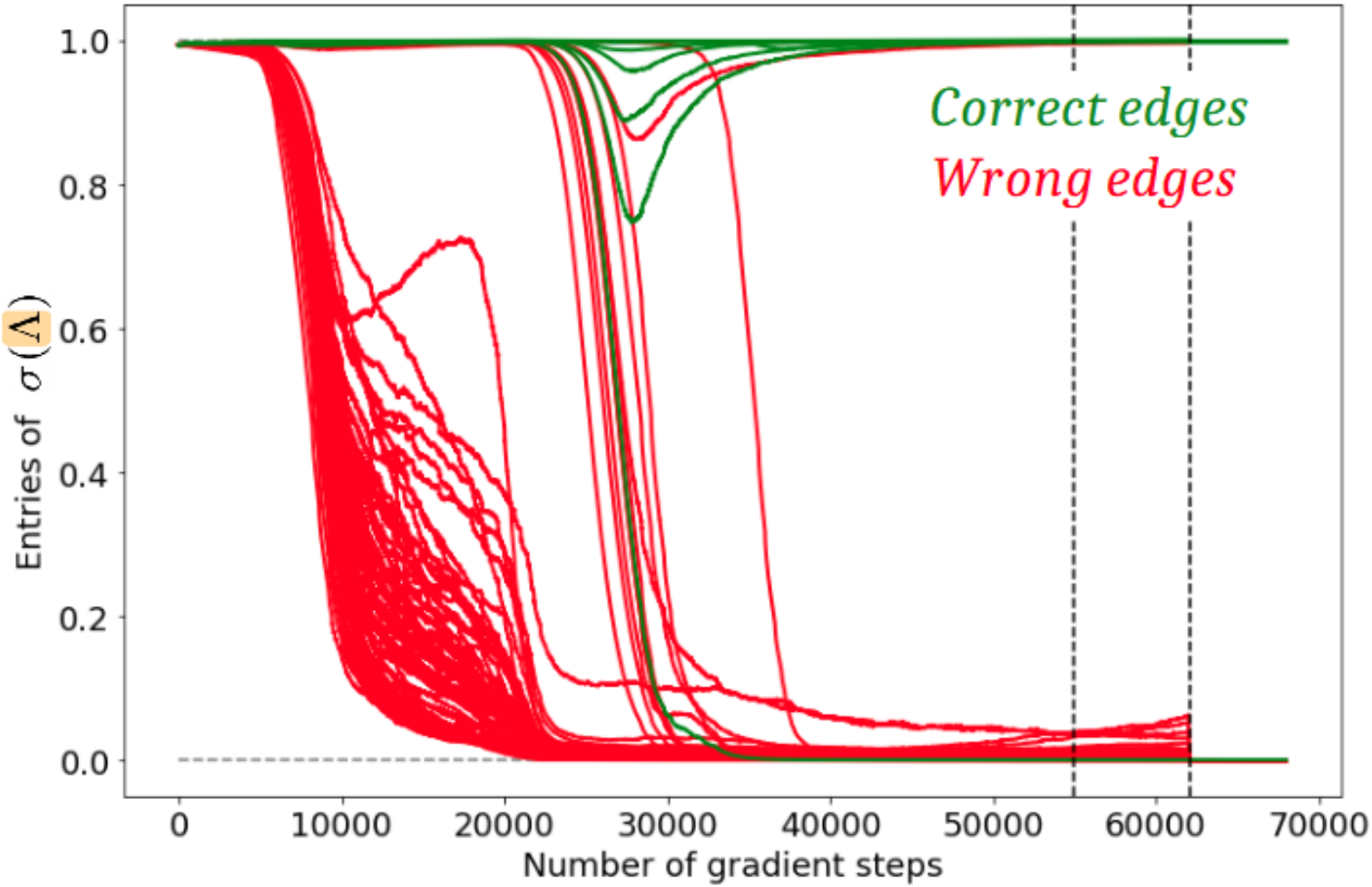
DCDI: optimization & gradient estimation

- Optimize jointly Λ and ϕ (NN parameters)

$$\max_{\phi,\Lambda} \mathop{\mathbb{E}}_{G\sim\sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X\sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda ||G||_0 \right] \text{ s.t. } \underbrace{\text{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

- **Optimization:** Augmented Lagrangian Method + RMSprop (as in Lachapelle et al. [2020])
- **Discrete sampling:** gradient w.r.t. Λ estimated via Gumbel-Softmax Straight-Through estimator [Jang et al., 2017; Maddison et al., 2017].
- Masks and/or the Gumbel-Softmax estimator were used before in causal discovery e.g., Kalainathan et al. [2018]; Ng et al. [2019]; Bengio et al. [2019]; Ke et al. [2019]

Result: structure learning via continuous optimization



Optimizing the objective gradually prunes anti-causal edges from the graph

DCDI: interventions with **unknown targets**

- Until now we assumed that I^* was known, i.e. we knew which variables were targeted.

- What if we don't? (e.g., as in Ke et al. [2019])

$$\mathcal{S}(G, \underbrace{I}_{\substack{\text{Intervention matrix} \\ \text{is learned}}}) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I, \phi) - \lambda \|G\|_0 - \underbrace{\lambda_I \|I\|_0}_{\substack{\text{Additional sparsity} \\ \text{regularizer}}}$$

- **Learning:**
  - Can do the same relaxation $I_{kj} \sim \text{Bernoulli}(\sigma(\beta_{kj}))$.
  - Optimize jointly for ϕ, Λ and β.

- **Theory:** We showed the same guarantee holds for this score!

DCDI: interventions with **unknown targets**

- Until now we assumed that I^* was known, i.e. we knew which variables were targeted.

- What if we don't? (e.g., as in Ke et al. [2019])

$$\mathcal{S}(G, \underbrace{I}_{\text{Intervention matrix is learned}}) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I, \phi) - \lambda \|G\|_0 - \underbrace{\lambda_I \|I\|_0}_{\text{Additional sparsity regularizer}}$$

- **Learning:**
  - Can do the same relaxation $I_{kj} \sim \text{Bernoulli}(\sigma(\beta_{kj}))$.
  - Optimize jointly for ϕ, Λ and β.

- **Theory:** We showed the same guarantee holds for this score!

DCDI: interventions with **unknown targets**

- Until now we assumed that I^* was known, i.e. we knew which variables were targeted.

- What if we don't? (e.g., as in Ke et al. [2019]) Learn it!

$$\mathcal{S}(G, \underbrace{I}_{\substack{\text{Intervention matrix} \\ \text{is learned}}}) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I, \phi) - \lambda \|G\|_0 - \underbrace{\lambda_I \|I\|_0}_{\substack{\text{Additional sparsity} \\ \text{regularizer}}}$$

- **Learning:**
  - Can do the same relaxation $I_{kj} \sim \text{Bernoulli}(\sigma(\beta_{kj}))$.
  - Optimize jointly for ϕ, Λ and β.

- **Theory:** We showed the same guarantee holds for this score!

DCDI: interventions with **unknown targets**

- Until now we assumed that I^* was known, i.e. we knew which variables were targeted.

- What if we don't? (e.g., as in Ke et al. [2019]) Learn it!

$$\mathcal{S}(G, \underbrace{I}_{\text{Intervention matrix is learned}}) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I, \phi) - \lambda \|G\|_0 - \underbrace{\lambda_I \|I\|_0}_{\text{Additional sparsity regularizer}}$$

- **Learning:**
  - Can do the same relaxation $I_{kj} \sim \text{Bernoulli}(\sigma(\beta_{kj}))$.
  - Optimize jointly for ϕ, Λ and β.

- **Theory:** We showed the same guarantee holds for this score!

DCDI: interventions with **unknown targets**

- Until now we assumed that I^* was known, i.e. we knew which variables were targeted.

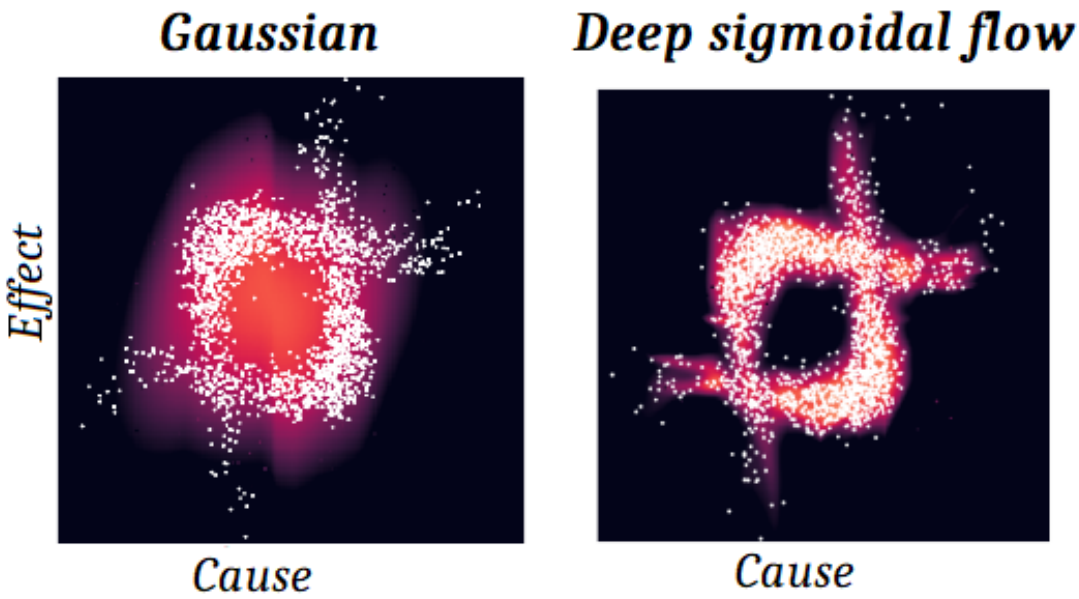
- What if we don't? (e.g., as in Ke et al. [2019]) Learn it!

$$\mathcal{S}(G, \underbrace{I}_{\text{Intervention matrix is learned}}) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I, \phi) - \lambda \|G\|_0 - \underbrace{\lambda_I \|I\|_0}_{\text{Additional sparsity regularizer}}$$

- **Learning:**
  - Can do the same relaxation $I_{kj} \sim \text{Bernoulli}(\sigma(\beta_{kj}))$.
  - Optimize jointly for ϕ, Λ and β.

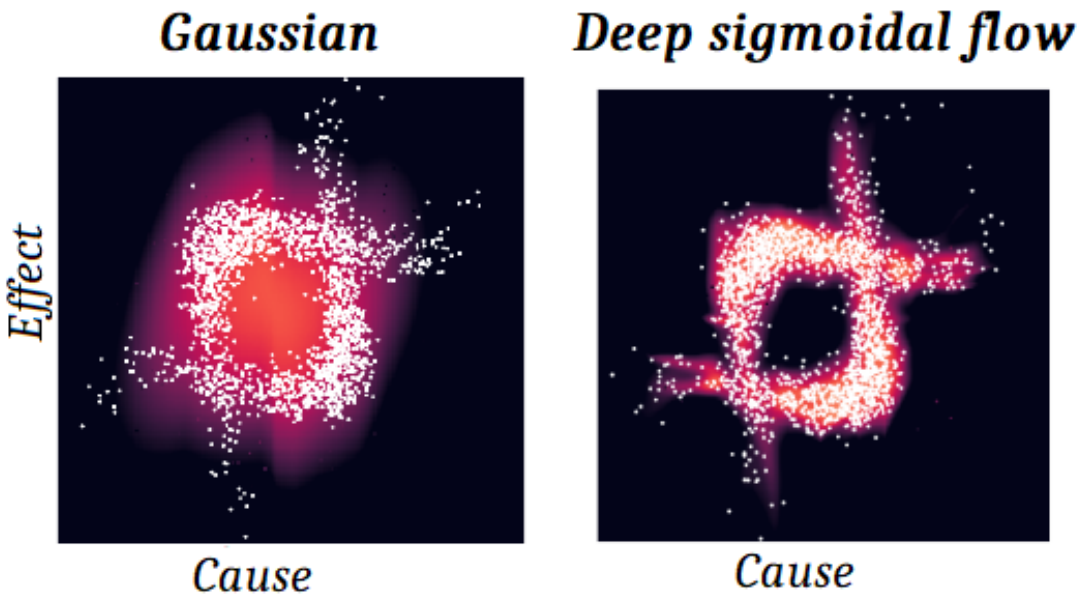
- **Theory:** We showed the same guarantee holds for this score!

DCDI: choice of density function $\tilde{f}$



- **Gaussian:** corresponds to a non-linear + additive noise assumption on the functional form of causal mechanisms
  - Identification guaranteed if it actually holds in the distribution [Peters et al., 2014]
- **Deep sigmoidal flow**: a type of normalizing flow that was shown to be a universal density approximator [Huang et al., 2018]
  - No assumption on functional forms
  - Identification guaranteed by our Thm 1 (with enough interventions)

DCDI: choice of density function $\tilde{f}$



- **Gaussian:** corresponds to a non-linear + additive noise assumption on the functional form of causal mechanisms
  - Identification guaranteed if it actually holds in the distribution [Peters et al., 2014]
- **Deep sigmoidal flow**: a type of normalizing flow that was shown to be a universal density approximator [Huang et al., 2018]
  - No assumption on functional forms
  - Identification guaranteed by our Thm 1 (with enough interventions)

Results – Structural Hamming Distance (lower is better)

DCDI-G = DCDI with Gaussian density
DCDI-DSF = DCDI with deep sigmoidal flow

ANM = nonlinear with additive noise
NN = nonlinear (no additive noise)
e = average number of parents

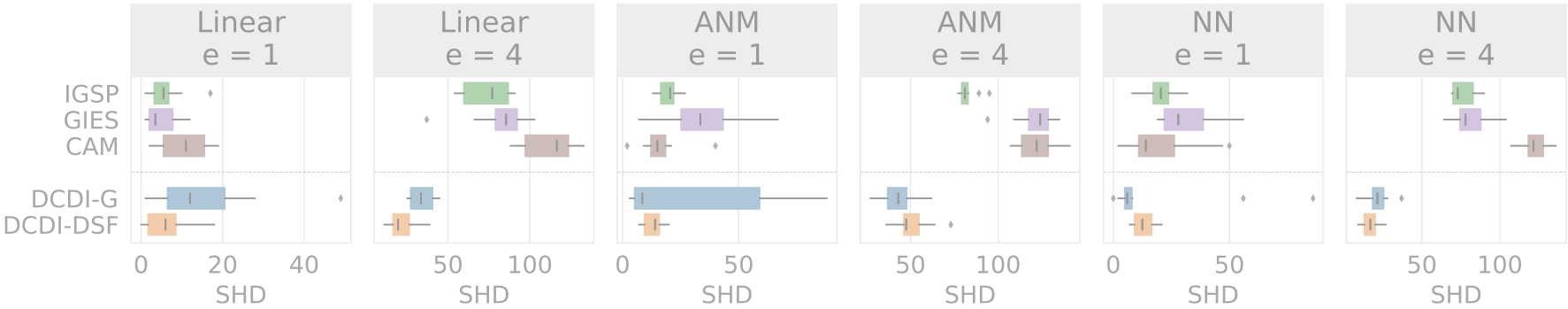


Figure: **Known target** interventions (20 nodes)

Figure: **Unknown target** interventions (20 nodes)

Results – Structural Hamming Distance (lower is better)

DCDI-G = DCDI with Gaussian density
DCDI-DSF = DCDI with deep sigmoidal flow

ANM = nonlinear with additive noise
NN = nonlinear (no additive noise)
e = average number of parents

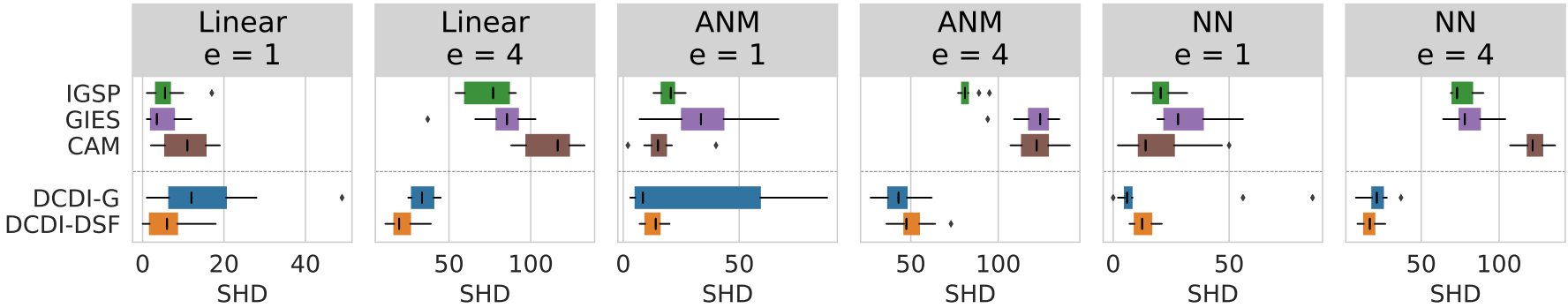


Figure: **Known target** interventions (20 nodes)

Figure: **Unknown target** interventions (20 nodes)

Results – Structural Hamming Distance (lower is better)

DCDI-G = DCDI with Gaussian density
DCDI-DSF = DCDI with deep sigmoidal flow

ANM = nonlinear with additive noise
NN = nonlinear (no additive noise)
e = average number of parents

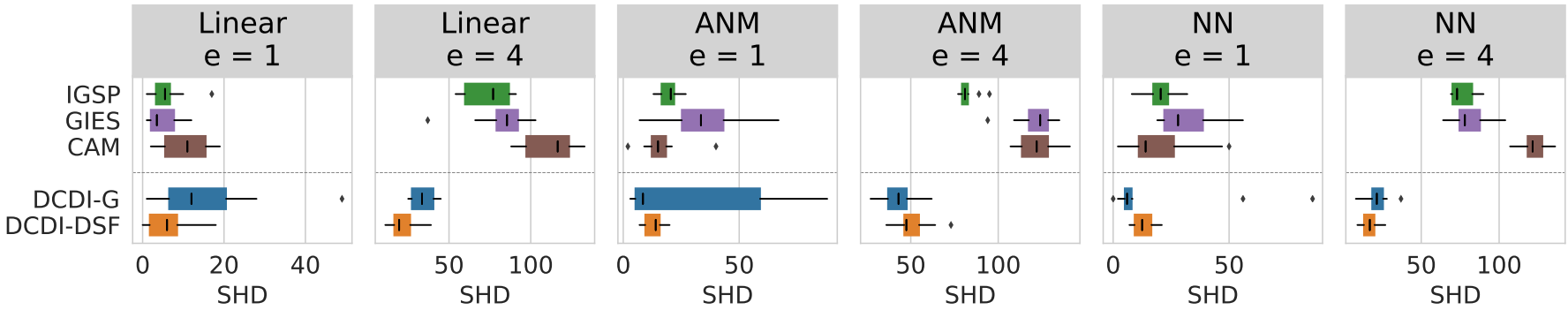


Figure: **Known target** interventions (20 nodes)

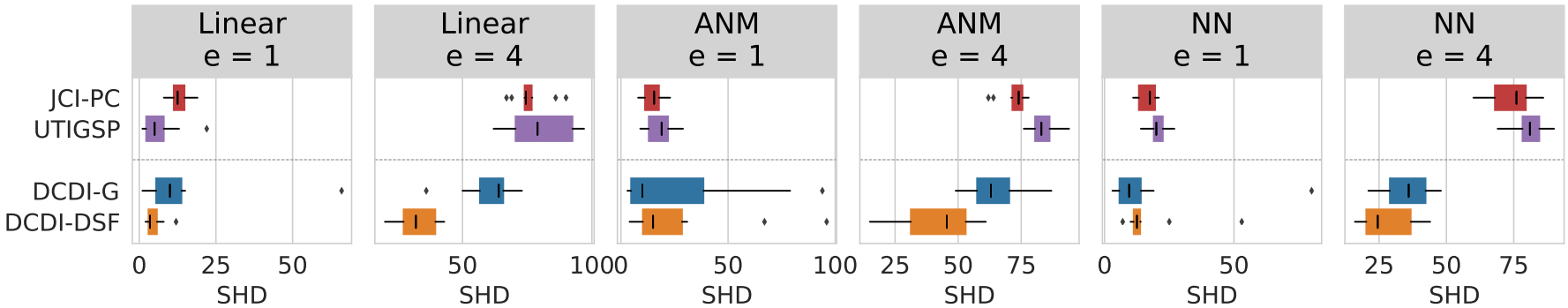


Figure: **Unknown target** interventions (20 nodes)

Conclusions and Future Directions

Conclusion & Future Work

We proposed DCDI, a causal discovery algorithm that:

- is **theoretically grounded**
- supports **perfect**, **imperfect** and **unknown-target** interventions
- **scales well with sample size** (compared to methods using kernel-based independence tests)
- achieves **state-of-the-art** performance, especially on denser graphs

Future work:

- **More extensive evaluation:** beyond synthetic data, violate assumptions [Gentzel et al., 2019]
- **Relax causal sufficiency:** allow for hidden confounders [Bhattacharya et al., 2020]
- **Time series:** non-stationnarities as imperfect interventions [Pamfil et al., 2020]
- **Learning variable representations:** not being agnostic to the nature of variables

Conclusion & Future Work

We proposed DCDI, a causal discovery algorithm that:

- is **theoretically grounded**
- supports **perfect**, **imperfect** and **unknown-target** interventions
- **scales well with sample size** (compared to methods using kernel-based independence tests)
- achieves **state-of-the-art** performance, especially on denser graphs

Future work:

- **More extensive evaluation:** beyond synthetic data, violate assumptions [Gentzel et al., 2019]
- **Relax causal sufficiency:** allow for hidden confounders [Bhattacharya et al., 2020]
- **Time series:** non-stationnarities as imperfect interventions [Pamfil et al., 2020]
- **Learning variable representations:** not being agnostic to the nature of variables

Conclusion & Future Work

We proposed DCDI, a causal discovery algorithm that:

- is **theoretically grounded**
- supports **perfect**, **imperfect** and **unknown-target** interventions
- **scales well with sample size** (compared to methods using kernel-based independence tests)
- achieves **state-of-the-art** performance, especially on denser graphs

Future work:

- **More extensive evaluation:** beyond synthetic data, violate assumptions [Gentzel et al., 2019]
- **Relax causal sufficiency:** allow for hidden confounders [Bhattacharya et al., 2020]
- **Time series:** non-stationnarities as imperfect interventions [Pamfil et al., 2020]
- **Learning variable representations:** not being agnostic to the nature of variables

Conclusion & Future Work

We proposed DCDI, a causal discovery algorithm that:

- is **theoretically grounded**
- supports **perfect**, **imperfect** and **unknown-target** interventions
- **scales well with sample size** (compared to methods using kernel-based independence tests)
- achieves **state-of-the-art** performance, especially on denser graphs

Future work:

- **More extensive evaluation:** beyond synthetic data, violate assumptions [Gentzel et al., 2019]
- **Relax causal sufficiency:** allow for hidden confounders [Bhattacharya et al., 2020]
- **Time series:** non-stationnarities as imperfect interventions [Pamfil et al., 2020]
- **Learning variable representations:** not being agnostic to the nature of variables

Conclusion & Future Work

We proposed DCDI, a causal discovery algorithm that:

- is **theoretically grounded**
- supports **perfect**, **imperfect** and **unknown-target** interventions
- **scales well with sample size** (compared to methods using kernel-based independence tests)
- achieves **state-of-the-art** performance, especially on denser graphs

Future work:

- **More extensive evaluation:** beyond synthetic data, violate assumptions [Gentzel et al., 2019]
- **Relax causal sufficiency:** allow for hidden confounders [Bhattacharya et al., 2020]
- **Time series:** non-stationnarities as imperfect interventions [Pamfil et al., 2020]
- **Learning variable representations:** not being agnostic to the nature of variables

Conclusion & Future Work

We proposed DCDI, a causal discovery algorithm that:

- is **theoretically grounded**
- supports **perfect**, **imperfect** and **unknown-target** interventions
- **scales well with sample size** (compared to methods using kernel-based independence tests)
- achieves **state-of-the-art** performance, especially on denser graphs

Future work:

- **More extensive evaluation:** beyond synthetic data, violate assumptions [Gentzel et al., 2019]
- **Relax causal sufficiency:** allow for hidden confounders [Bhattacharya et al., 2020]
- **Time series:** non-stationnarities as imperfect interventions [Pamfil et al., 2020]
- **Learning variable representations:** not being agnostic to the nature of variables

Conclusion & Future Work

We proposed DCDI, a causal discovery algorithm that:

- is **theoretically grounded**
- supports **perfect**, **imperfect** and **unknown-target** interventions
- **scales well with sample size** (compared to methods using kernel-based independence tests)
- achieves **state-of-the-art** performance, especially on denser graphs

Future work:

- **More extensive evaluation:** beyond synthetic data, violate assumptions [Gentzel et al., 2019]
- **Relax causal sufficiency:** allow for hidden confounders [Bhattacharya et al., 2020]
- **Time series:** non-stationnarities as imperfect interventions [Pamfil et al., 2020]
- **Learning variable representations:** not being agnostic to the nature of variables

Conclusion & Future Work

We proposed DCDI, a causal discovery algorithm that:

- is **theoretically grounded**
- supports **perfect**, **imperfect** and **unknown-target** interventions
- **scales well with sample size** (compared to methods using kernel-based independence tests)
- achieves **state-of-the-art** performance, especially on denser graphs

Future work:

- **More extensive evaluation:** beyond synthetic data, violate assumptions [Gentzel et al., 2019]
- **Relax causal sufficiency:** allow for hidden confounders [Bhattacharya et al., 2020]
- **Time series:** non-stationnarities as imperfect interventions [Pamfil et al., 2020]
- **Learning variable representations:** not being agnostic to the nature of variables

Thank you!


Philippe Brouillard *[1]


Sébastien Lachapelle *[1]


Alexandre Lacoste [2]


Simon Lacoste-Julien [1]


Alexandre Drouin [2]

[1] Mila & DIRO, Université de Montréal [2] Element AI / ServiceNow * Equal contribution

https://github.com/slachapelle/dcdi

alexandre.drouin@servicenow.com @alexandredrouin

References

Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., & Pal, C. (2019). A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*.

Bhattacharya, R., Nagarajan, T., Malinsky, D., & Shpitser, I. (2020). Differentiable causal discovery under unmeasured confounding. *arXiv preprint arXiv:2010.06978*.

Chickering, D. (2003). Optimal structure identification with greedy search. *Journal of Machine Learning Research*.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., & Regev, A. (2016). Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7), 1853–1866.e17.

Eberhardt, F., Glymour, C., & Scheines, R. (2005). On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05 (pp. 178184). Arlington, Virginia, USA: AUAI Press.

Gentzel, A., Garant, D., & Jensen, D. (2019). The case for evaluating causal models using interventional measures and empirical data. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (pp. 11717–11727).

Huang, C.-W., Krueger, D., Lacoste, A., & Courville, A. (2018). Neural autoregressive flows.

Huang, Y. & Valtorta, M. (2006). Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06 (pp. 217224). Arlington, Virginia, USA: AUAI Press.

Jang, E., Gu, S., & Poole, B. (2017). Categorical reparameterization with gumbel-softmax. *Proceedings of the 34th International Conference on Machine Learning*.

Julious, S. A. & Mullee, M. A. (1994). Confounding and simpson's paradox. *Bmj*, 309(6967), 1480–1481.

Kalainathan, D., Goudet, O., Guyon, I., Lopez-Paz, D., & Sebag, M. (2018). Sam: Structural agnostic model, causal discovery and penalized adversarial learning. *arXiv preprint arXiv:1803.04929*.

Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Pal, C., & Bengio, Y. (2019). Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*.

Koller, D. & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. MIT Press.

Lachapelle, S., Brouillard, P., Deleu, T., & Lacoste-Julien, S. (2020). Gradient-based neural DAG learning. In *Proceedings of the 8th International Conference on Learning Representations*.

Maddison, C. J., Mnih, A., & Teh, Y. W. (2017). The concrete distribution: A continuous relaxation of discrete random variables. *Proceedings of the 34th International Conference on Machine Learning*.

Ng, I., Fang, Z., Zhu, S., Chen, Z., & Wang, J. (2019). Masked gradient-based causal structure learning. *arXiv preprint arXiv:1910.08527*.

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., & Aragam, B. (2020). Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics* (pp. 1595–1605).: PMLR.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of Causal Inference - Foundations and Learning Algorithms*. MIT Press.

Peters, J., M. Mooij, J., Janzing, D., & Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*.

Shpitser, I. & Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21 (pp. 1219).: Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Verma, T. & Pearl, J. (1991). *Equivalence and synthesis of causal models*. UCLA, Computer Science Department.

Yang, K. D., Katcoff, A., & Uhler, C. (2018). Characterizing and learning equivalence classes of causal DAGs under interventions. *Proceedings of the 35th International Conference on Machine Learning*.

Zheng, X., Aragam, B., Ravikumar, P., & Xing, E. (2018). Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems 31*.