Differentiable Causal Discovery from Interventional Data
(DCDI)





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
* Equal contribution

Contributions

Differentiable Causal Discovery with Interventions (DCDI) is causal discovery algorithm that

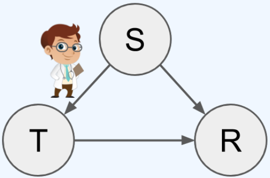
- can leverage **perfect**, **imperfect** and **unknown-target** interventions;
- relies on **continuous-constrained optimization** and **neural networks**;
- does not make strong parametric assumptions about the causal mechanisms, thanks to expressive **normalizing flows**;
- is **theoretically grounded**;
- and **compares favorably** to SOTA methods.

Causal graphical models (CGM)

- Random vector $X = (X_1, ..., X_d)$
- Let $\mathcal{G}$ be a **directed acyclic graph** (DAG)
- $\mathcal{G}$ describes causal relationships between variables.

Example: Kidney stone treatment

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



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

Perfect and Imperfect Interventions

- CGM can model **interventions**, i.e. a localized change in a distribution.

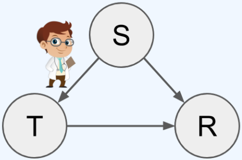
Mila Université de Montréal
ELEMENT AI

- CGM can model **interventions**, i.e. a localized change in a distribution.

Intervening on the treatment T

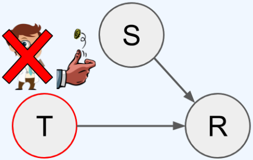
T = Treatment $\in \{A, B\}$
S = Stone size $\in$ {small, large}
R = Patient recovered $\in \{0, 1\}$

Observations



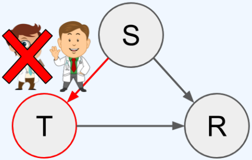
$p(S)p(T \mid S)p(R \mid S, T)$

Perfect intervention



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

Imperfect intervention



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

Mila Université de Montréal ELEMENT AI

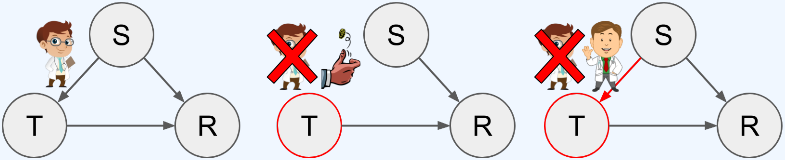
Perfect and Imperfect Interventions

- CGM can model **interventions**, i.e. a localized change in a distribution.

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 | **Perfect intervention** | **Imperfect intervention**

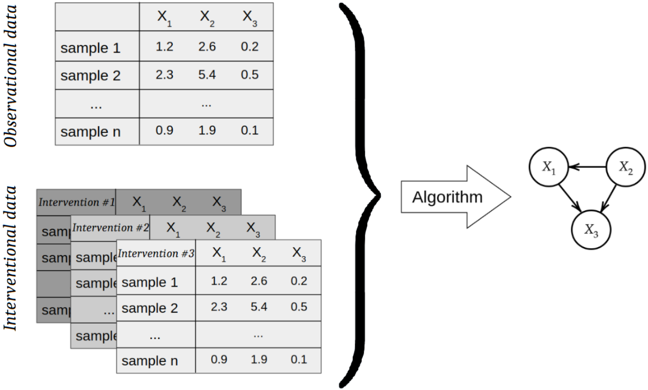


$p(S)p(T \mid S)p(R \mid S, T)$ | $p(S)\tilde{p}(T)p(R \mid S, T)$ | $p(S)\tilde{p}(T \mid S)p(R \mid S, T)$

The ability to model interventions is **crucial to predict the effect of actions/policies** and **requires the causal graph**.

Causal discovery from interventions

But the causal graph might be **unknown**...
Causal discovery = learn the causal graph!



Problem setting and notation

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

Mila Université de Montréal
ELEMENT AI

Problem setting and notation

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

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

Mila Université de Montréal
ELEMENT AI

Problem setting and notation

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

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

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

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

Problem setting and notation

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

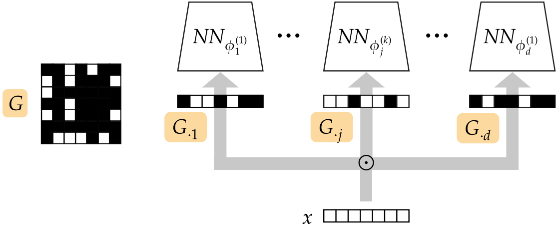
$$\text{Causal DAG} = G = \underbrace{\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}}_{\text{Adjacency matrix}} \in \{0,1\}^{d \times d}$$

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

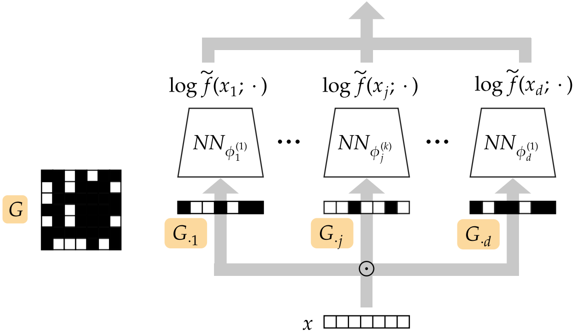
$$\underbrace{I = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}}_{\text{Intervention matrix}} \in \{0,1\}^{K \times d}$$

- $G^* =$ ground truth causal graph.
- $I^* =$ ground truth intervention matrix.

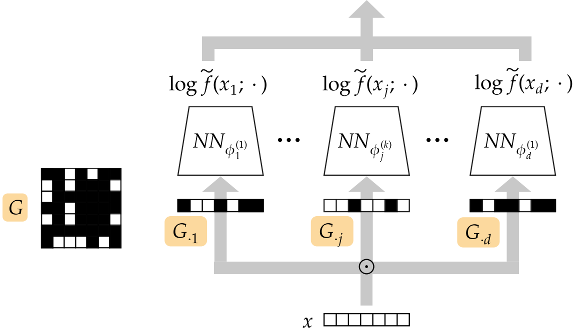
DCDI: The model



DCDI: The model

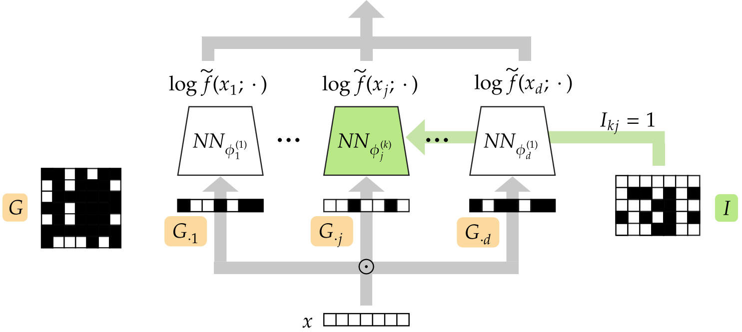


$$f^{(k)}(x; G, I, \phi) := \prod_{j=1}^{d} \tilde{f}(x_j; \mathrm{NN}(G_j \odot x; \overbrace{\phi_j^{(1)}}^{\text{Observational parameter}}))^{1-I_{kj}} \tilde{f}(x_j; \mathrm{NN}(G_j \odot x; \overbrace{\phi_j^{(k)}}^{\text{Interventional parameter}}))^{I_{kj}}$$



DCDI: The model

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



DCDI: The score (discrete form)

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

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

- Here, we assume I^* is **known** (we relax this assumption later!).
- We will see later how to relax to a continuous constrained problem.

Mila Université de Montréal

ELEMENT AI

DCDI: Theoretical justification

- $G^* =$ ground-truth DAG
- $I^* =$ ground-truth intervention matrix

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

Theorem (Identification via score maximization)

Suppose $I^*_{1,:} = 0$. 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

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

Mila Université de Montréal

ELEMENT AI

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

DCDI: Theoretical justification

- $G^* =$ ground-truth DAG
- $I^* =$ ground-truth intervention matrix

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

Theorem (Identification via score maximization)

Suppose $I^*_{1,:} = 0$. 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

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

More general result

Without the first assumption, we can identify the I^*-Markov equivalence class[1] of G^*.

Université de Montréal

ELEMENT AI

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

DCDI: Continuous-constrained formulation

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

$\Downarrow$

Relaxation where $G_{ij} \sim$ Bernoulli$(\sigma(\Lambda_{ij}))$,

with $\sigma(\cdot) :=$ sigmoid function

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

DCDI: Continuous-constrained formulation

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

Relaxation where $G_{ij} \sim$ 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

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

[Zheng et al., 2018]

DCDI: Optimization & Gradient estimation

- Optimize jointly Λ and ϕ (NN parameters)

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

- Optimized with **RMSprop** + **augmented Lagrangian method**.

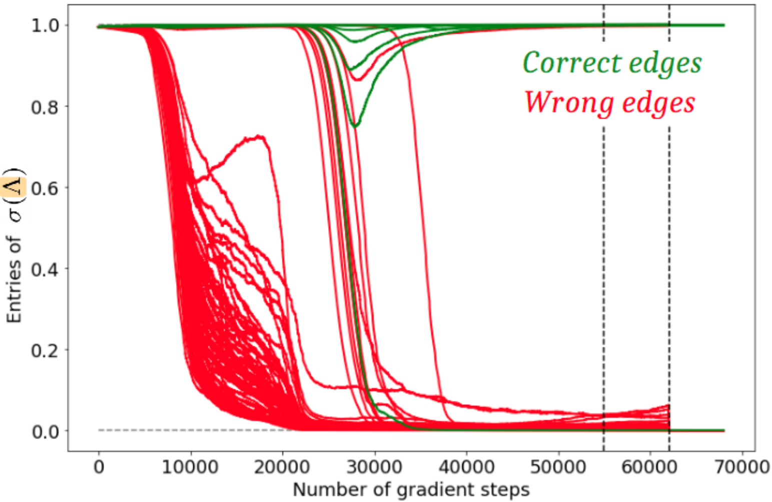
DCDI: Optimization & Gradient estimation

- Optimize jointly Λ and ϕ (NN parameters)

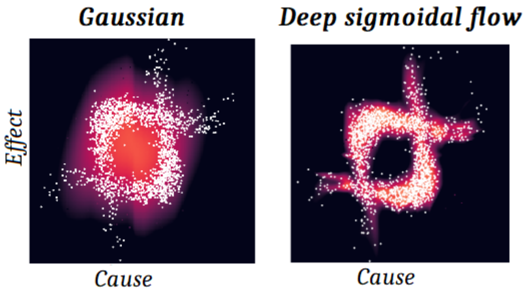
$$\max_{\phi,\Lambda} \mathop{\mathbb{E}}_{G\sim\sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X\sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda ||G||_0 \right] \text{ s.t. } \underbrace{\mathrm{Tr} e^{\sigma(\Lambda)} - d = 0}_{\text{Acyclicity constraint}}$$

- Optimized with **RMSprop** + **augmented Lagrangian method**.
- 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:
[Kalainathan et al., 2018, Ng et al., 2019, Bengio et al., 2019, Ke et al., 2019]

DCDI: Optimization



Choice of density function $\tilde{f}$



- Deep sigmoidal flow [Huang et al., 2018] = a specific kind of **normalizing flow**.
- Gaussian fails to recover the causal direction while the normalizing flow can (Not visible from the plot).

Mila Université de Montréal
ELEMENT AI

Support for interventions with **unknown targets**

- Up to now we assumed I^* is known, i.e. we knew which variables were targeted.
- What if it is **unknown**? Learn it!

Support for interventions with **unknown targets**

- Up to now we assumed I^* is known, i.e. we knew which variables were targeted.
- What if it is **unknown**? Learn it!

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

Support for interventions with **unknown targets**

- Up to now we assumed I^* is known, i.e. we knew which variables were targeted.
- What if it is **unknown**? Learn it!

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

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

Support for interventions with **unknown targets**

- Up to now we assumed I^* is known, i.e. we knew which variables were targeted.
- What if it is **unknown**? Learn it!

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

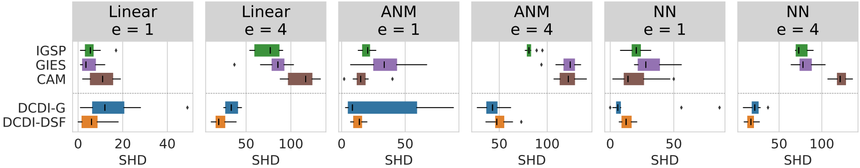
- We showed the same **theoretical guarantee** holds for this score!
- Can do the same relaxation $I_{kj} \sim \text{Bernoulli}(\sigma(\beta_{kj}))$.
- Optimize jointly for ϕ, Λ and β.

Mila Université de Montréal
ELEMENT AI

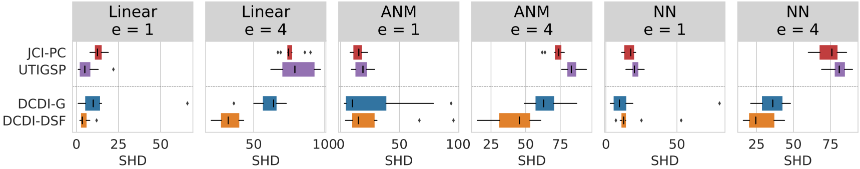
Experiment summary (lower is better)

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



Known target interventions (20 nodes)



Unknown target interventions (20 nodes)

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; and
- **works better for denser graphs** compared to other greedy search methods.

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; and
- **works better for denser graphs** compared to other greedy search methods.

Future work:

- Relax **causal sufficiency**, i.e. allow for hidden confounders;
- **Scaling up to larger graphs** (> 100 nodes):
  The matrix exponential from the acyclicity constraint costs $\mathcal{O}(d^3)$.

Learn more

If you want to know more about DCDI:

- Check our paper
- Check our github repo: https://github.com/slachapelle/dcdi
- Come talk to us!

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.

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

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

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.

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.

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.*