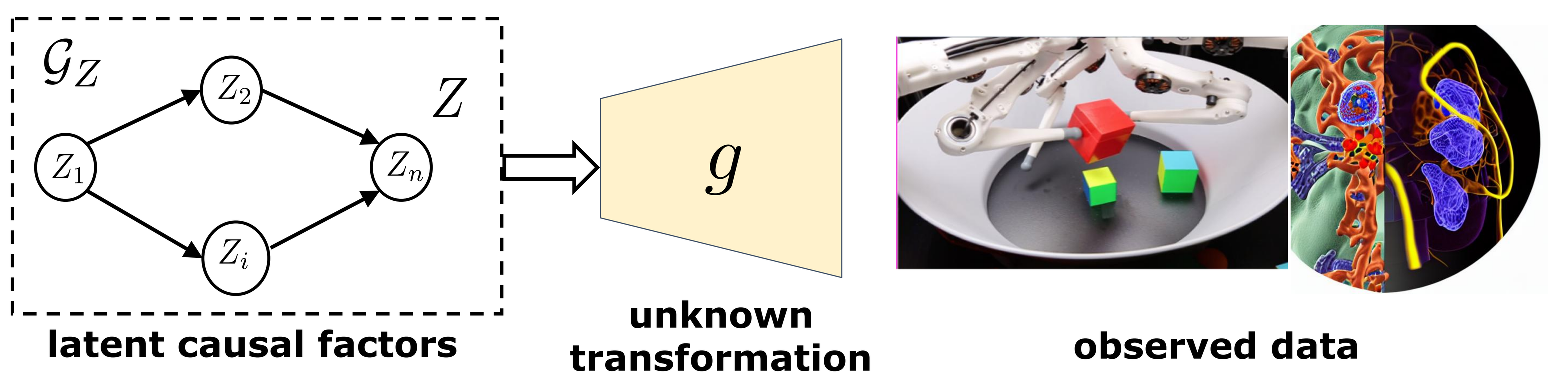
Sample Complexity of Interventional Causal Representation Learning

Emre Acartürk[1] Burak Varıcı[2] Karthikeyan Shanmugam[3] Ali Tajer[1]

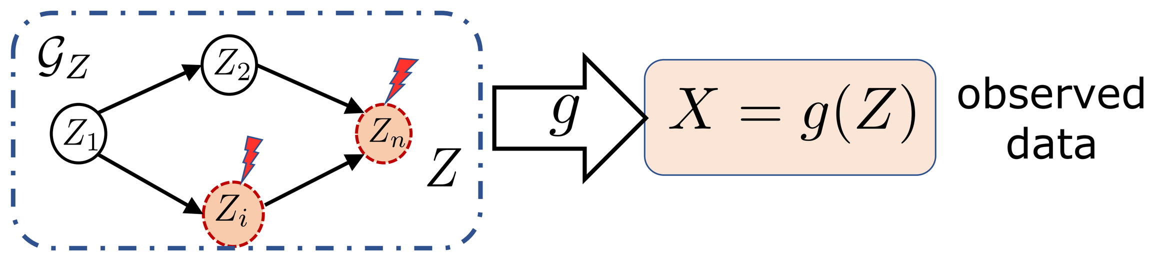
[1]Rensselaer Polytechnic Institute [2]Carnegie Mellon University [3]Google DeepMind India

Causal Representation Learning



Generic goal: Invert the unknown transformation to recover

1) latent representation and **2) the latent causal structure**



- **Identifiability:** (im)possibility of uniquely* recovering Z and $\mathcal{G}_Z$
- **Achievability:** provably correct and scalable algorithms

existing literature: **asymptotic** guarantees (infinite samples)

What are finite-sample guarantees?

Problem Setting

- **Linear CRL:** Transformation g is linear, i.e., $X = \mathbf{G} \cdot Z$
- **Single-node soft interventions:** Most general form

$$p_Z^m(z) = q_i(z_i \mid z_{\mathrm{pa}(i)}) \prod_{j \neq i} p_j(z_j \mid z_{\mathrm{pa}(j)})$$

- **Finite sample data:** N samples of X per environment

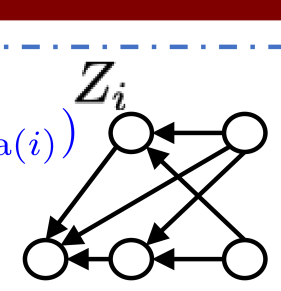
Identifiability Objective

(ϵ, δ)**–PAC identifiability:** The same infinite-sample identifiability guarantees with probability at least $(1 - \delta)$

Infinite-sample guarantees:

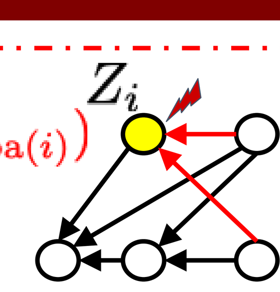
- $\hat{\mathcal{G}}_Z$ is equal to the transitive closure of $\mathcal{G}_Z$
- $\hat{Z}_i$ is a linear function of $Z_i \cup \{Z_j \ : \ j \in \mathrm{pa}(i)\}$

Main Tool: Score Differences



$p_i(z_i \mid z_{\mathrm{pa}(i)})$

$$p_Z^m(z) = p_i(z_i \mid z_{\mathrm{pa}(i)}) \prod_{j \neq i} p_j(z_j \mid z_{\mathrm{pa}(j)})$$



$q_i(z_i \mid z_{\mathrm{pa}(i)})$

$$p_Z^m(z) = q_i(z_i \mid z_{\mathrm{pa}(i)}) \prod_{j \neq i} p_j(z_j \mid z_{\mathrm{pa}(j)})$$

$\log p_Z^0(z) - \log p_Z^m(z) = \log \frac{p_i}{q_i}(z_i | z_{\mathrm{pa}(i)})$: function of **only** z_i and $z_{\mathrm{pa}(i)}$

Define score function and score difference:

$$s_Z^m(z) \triangleq \nabla_z \log p_Z^m(z) \quad \text{and} \quad d_Z^m(z) \triangleq s_Z^m(z) - s_Z^0(z)$$

$$d_Z^m(z) = [0 \ \ 0 \ \ \boxed{\times} \ \ 0 \ \ \boxed{\times} \ \ 0 \ \ 0 \ \ \underset{i}{\boxed{\times}} \ \ 0]^\top$$

coordinates of parents of node i

Score function and difference can be defined for X too

$$s_X^m(x) \triangleq \nabla_x \log p_X^m(x) \quad \text{and} \quad d_X^m(x) \triangleq s_X^m(x) - s_X^0(x)$$

Observation space score differences are intimately related

$$d_X^m(x) = \left(\mathbf{G}^\dagger\right)^\top \cdot d_Z^m(z)$$

Both **inverse transform** and **latent graph** information are encoded in **observed score differences**.

Methodology

Core observation: Using the image/column spaces of $d_X^m(x)$ suffice to recover both!

Infinite-sample algorithm:

- Achieve identifiability using **only** column spaces of $d_X^m(x)$
- Check only matrix **rank** and subspace **orthogonality**

Finite-sample algorithm:

- Replace column space of $d_X^m(x)$ with the approximate column space of $\hat{d}_X^m(x)$
- Show, using enough samples, with high probability,

$$\text{rank}(d_X^m(x)) = \text{est. rank}(\hat{d}_X^m(x)) \ ,$$

similarly for approximate orthogonality.

Results

Consider a generic consistent score (difference) estimator, i.e.,

$$\mathbb{P}\left(\max_{m \in [n]} \left\| \hat{d}_X^m(x) - d_X^m(x) \right\|_2 > \epsilon \right) < \delta \ , \qquad \forall N \geq N(\epsilon, \delta) \ .$$

Under a mild regularity assumption on p_Z that ensures the effect of an intervention is distinct between Z_i and $Z_{\mathrm{pa}(i)}$,

Theorem (Sample complexity – general). For any consistent score difference estimator with sample complexity $N(\epsilon, \delta)$, we achieve (ϵ, δ)–PAC identifiability when

$$N \geq N\left(\min\left\{\epsilon \cdot \kappa, \epsilon_{\min}\right\}, \delta\right)$$

where κ and $\epsilon_{\min}$ are model constants.

Adopting a specific score estimator,

Theorem (Sample complexity – RKHS). Using a reproducing kernel Hilbert space-based score estimator [1], we achieve (ϵ, δ)–PAC identifiability when

$$N \geq C \cdot \left(\max\left\{\frac{1}{\epsilon}, c\right\}\right)^4 \cdot \left(\frac{1}{\delta}\right)^4$$

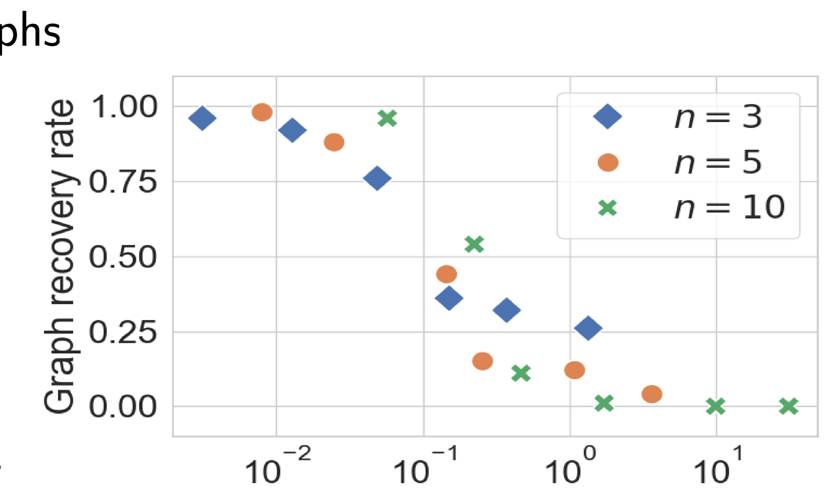
where κ and $\epsilon_{\min}$ are model constants.

The first complexity result for interventional CRL.

(Constants are all exactly specified)

Experiments

- Linear Gaussian SEMs, Erdős–Rényi random graphs (100 runs)
- Latent dimension $n \in \{3, 5, 10\}$, observed dimension $d \in \{n, 15\}$
- Number of samples $N \in \{10^{2.5}, 10^3, 10^{3.5}, 10^4, 10^{4.5}, 10^5\}$
- Plot rate of perfect graph recovery vs MSE of score estimator



Check out other score-based CRL work!

- **General transformations:** "*General identifiability and achievability for causal representation learning*". AISTATS 2024.
- **Single-node interventions** (base for this paper): "*Score-based causal representation: Linear and general transformations*". arXiv: 2402.00849
- **Multi-node interventions!** "Linear Causal Representation Learning from Unknown Multi-node Interventions". NeurIPS 2024

[1] Yuhao Zhou, Jiaxin Shi, and Jun Zhu. Nonparametric score estimators. ICML 2020

Contact: acarte@rpi.edu