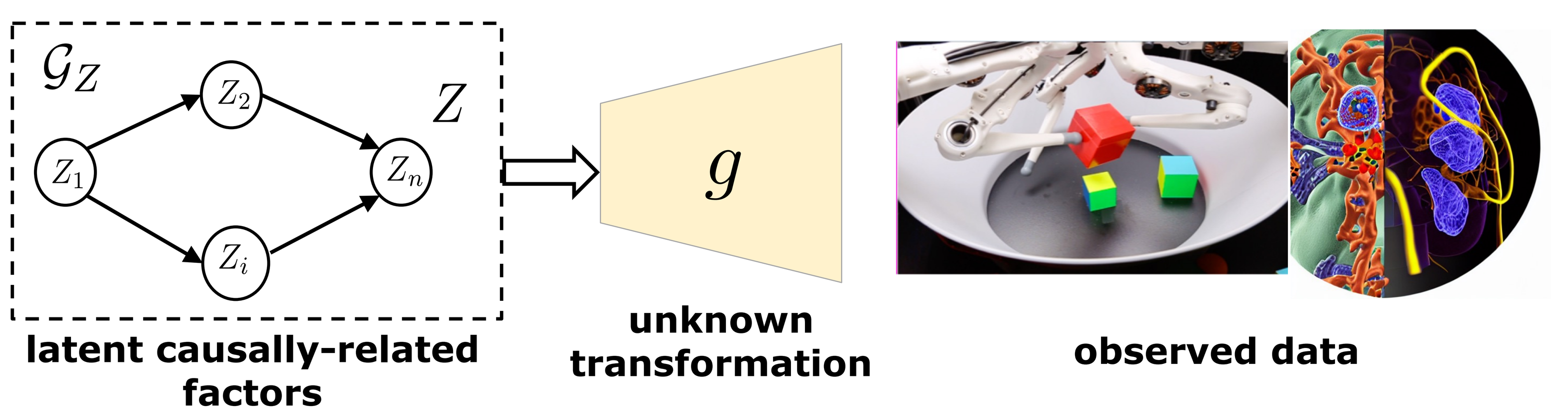
Linear Causal Representation Learning from Unknown Multi-node Interventions

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

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

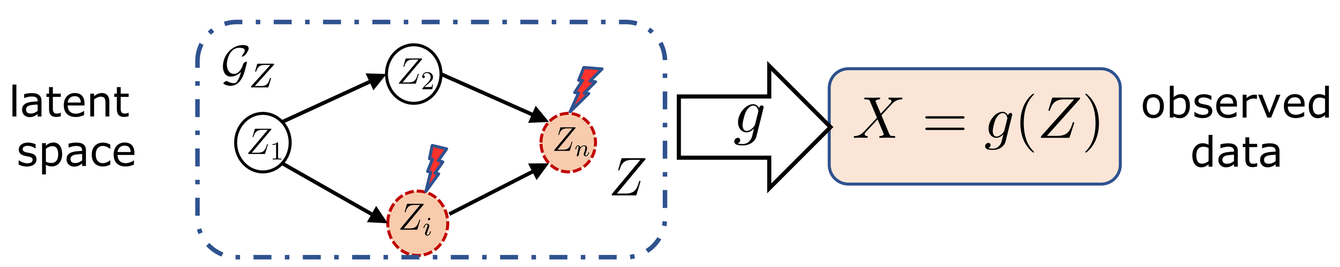
We learn **causal representations** using *unknown multi-node interventions* on latent space by leveraging score functions

Causal Representation Learning



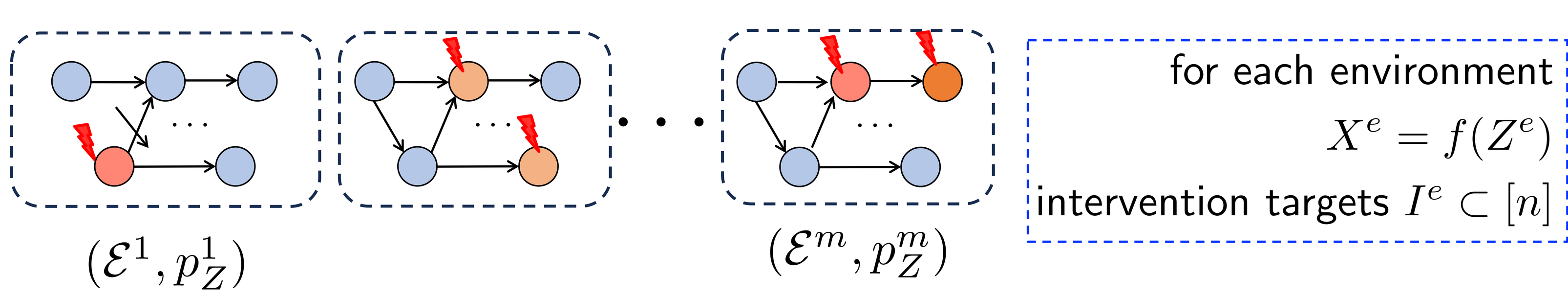
Generic goal: Inverting the unknown transformation to recover

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



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

Interventional CRL



- **Distribution level** info: Multiple datasets, *almost* unsupervised
- Distr. shifts: changes in causal mechanisms $p_i(z_i|z_{\mathrm{pa}(i)}) \to q_i(z_i|z_{\mathrm{pa}(i)})$
- Prior work: single-node interventions

this paper: UNKNOWN MULTI-NODE INTERVENTIONS

Multi-node interventions: $M \geq n$ environments
unknown interv. targets $I^m \subset [n]$, for $m \in [M]$

env. $\mathcal{E}^m$ with I^m : $\quad p^m(z) = \prod_{i \in I^m} q_i(z_i|z_{\mathrm{pa}(i)}) \prod_{i \notin I^m} p_i(z_i|z_{\mathrm{pa}(i)})$

Linear transform: $\quad X = \mathbf{G} \cdot Z$

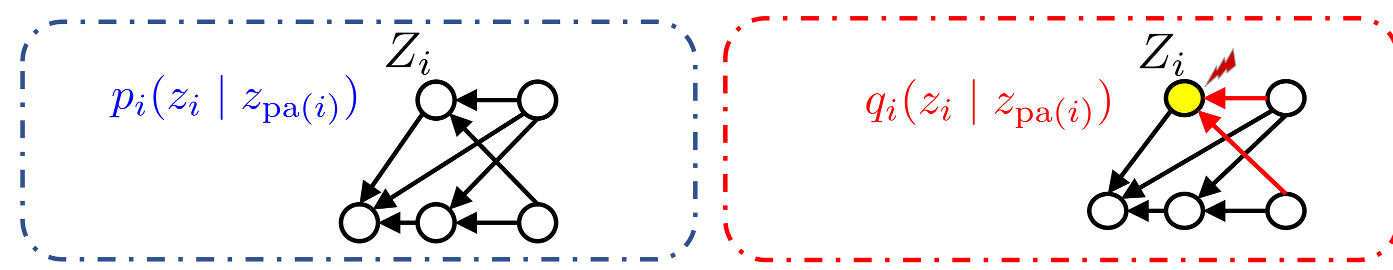
Score functions

Observational : $s(z) \triangleq \nabla_z \log p(z)$ and $s_X(x) \triangleq \nabla_x \log p_X(x)$

Interventional : $s^m(z) \triangleq \nabla_z \log p^m(z)$ and $s_X^m(x) \triangleq \nabla_x \log p_X^m(x)$

Property: $\quad s(z) = \mathbf{G}^\top \cdot s_X(x)$

Why score functions?



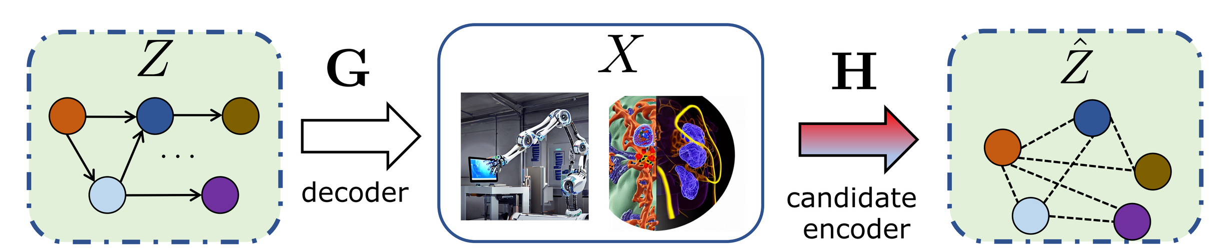
$$\log p(z) - \log p^m(z) = \log p_i(z_i|z_{\mathrm{pa}(i)}) - \log q_i(z_i|z_{\mathrm{pa}(i)})$$

function of only z_i and $z_{\mathrm{pa}(i)}$

$$s(z) - s^m(z) = \begin{bmatrix} 0 & 0 & \boxed{\times} & 0 & \boxed{\times} & 0 & 0 & \underset{i}{\boxed{\times}} & 0 \end{bmatrix}^\top$$

coordinates of parents of node i

Score functions contain all information about latent DAG



incorrect encoder $\rightarrow$ $\quad s_{\hat{Z}}(\hat{z}) - s_{\hat{Z}}^m(\hat{z})$ **not** a function of only z_i & $z_{\mathrm{pa}(i)}$

estimated score differences cannot be sparser than true score differences

estimate of inverse transform $\mathbf{G}^\dagger$ =
encoder $\mathbf{H}$ that minimizes score variations

Methodology

Challenges for multi-node interventions

- The intervention targets are fully unknown!
- Latent score differences are not sparse for multi-node:

$$s(z) - s^m(z) = \sum_{i \in I^m} \nabla_z \log \frac{p_i}{q_i}(z_i \mid z_{\mathrm{pa}(i)})$$

$$\left\| \mathbb{E}[s(z) - s^m(z)] \right\|_0 = \left| \cup_{i \in I^m} \mathrm{pa}(i) \cup \{i\} \right|$$

1. Combine multi-node interv. to create sparser interventions

- **Idea:** if intervention targets are diverse, reduce to single-int. problem
- Example: Given $I^1 = \{1\}, I^2 = \{1,3\}, I^3 = \{2,3\}$ and $I^0 = \emptyset$
- $s^2(z) - s^1(z)$ gives $\tilde{I}^3 = \{3\}$, $\quad s^3(z) - \tilde{s}^3(z)$ gives $\tilde{I}^2 = \{2\}$

2. How to do it with unknown intervention targets?

- Consider $s_X, s_X^1, \dots, s_X^n$. Iteratively search mixing vectors $\mathbf{w} \in \mathbb{N}^n$

$$\dim\Big(\mathrm{proj.image}\big(\sum_{i \in [n]} w_i \cdot (s_X - s_X^i)\big)\Big) = 1$$

- Why? Single-node root intervention, i-th row of $\mathbf{G}^\dagger = \mathrm{image}(s_X - s_X^i)$
- If score difference is not minimized, then dimension of the image > 1
- Encoder estimate: choose $\mathbf{H} \in \mathrm{image}(\Delta \mathbf{S}_X \cdot \mathbf{W}) \subset \mathbb{R}^{d \times n}$

Results

Theorem (soft): Using diverse, regular unknown **multi-node soft** interventions, we have **identifiability up to ancestors**:

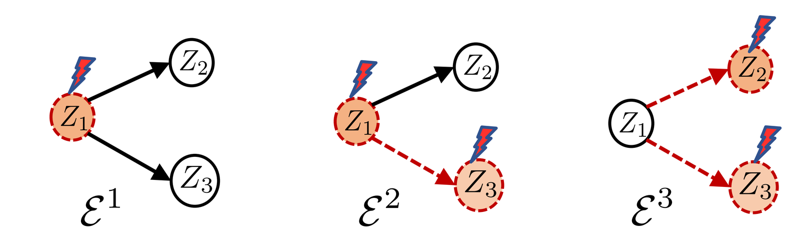
- $\hat{Z}_i$ is a linear function of $Z_i \cup \{Z_j : j \in \mathrm{ancestors}(i)\}$,
- $\hat{\mathcal{G}}_Z$ is transitive closure of $\mathcal{G}_Z$

Theorem (hard): Using diverse, regular unknown **multi-node hard** interventions and additive noise, we have **perfect identifiability**:

- $\hat{Z}_i = c_i \times Z_i$ for a constant scalar c_i, and $\hat{\mathcal{G}}_Z = \mathcal{G}_Z$

Same guarantees as single-node interventions!

Diverse: full-rank intervention matrix $\mathbf{D} \in \{0,1\}^{n \times M}$ with $\mathbf{D}_{i,m} = \mathbb{I}(i \in I^m)$



$$\mathbf{D} = \begin{array}{c} \begin{array}{ccc} \mathcal{E}^1 & \mathcal{E}^2 & \mathcal{E}^3 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \end{array} \begin{array}{c} -Z_1 \\ -Z_2 \\ -Z_3 \end{array}$$

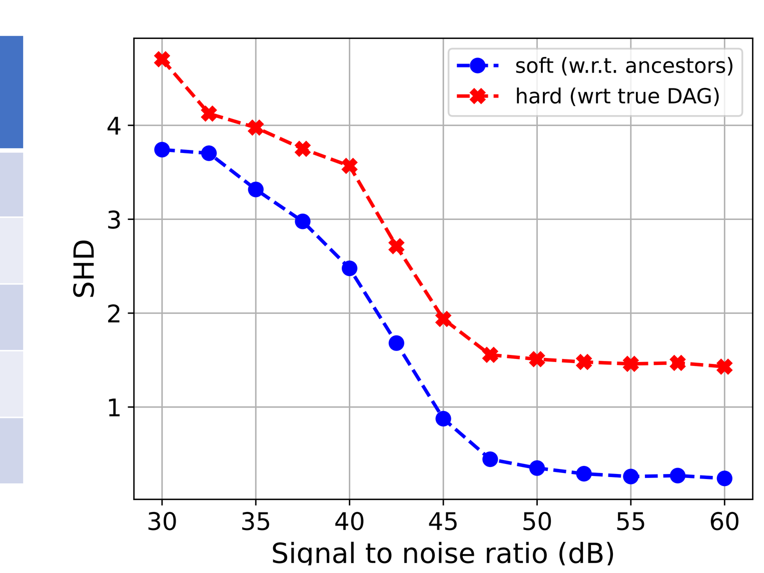
Regularity (informal): Effect of a multi-node intervention is not the same on the scores associated with different nodes.

Experiments

- Linear Gaussian SEMs with Erdős–Rényi random graphs (100 runs)
- Scores: $s_X(x) = -\Theta \cdot x$, estimate precision matrix Θ with 10^5 samples
- Sensitivity analysis for quadratic models (ground truth scores + noise)
- Structural Hamming distance (SHD) for latent graph (ideally 0)
- Mean correlation coefficient (MCC) for latent variables (ideally 1)

Latent dim.	Soft SHD	Soft MCC	Hard SHD	Hard MCC
4	0.77	**0.96**	0.66	**0.98**
5	1.93	**0.93**	1.80	**0.98**
6	3.39	**0.92**	3.05	**0.95**
7	4.62	**0.91**	6.12	**0.91**
8	8.26	**0.90**	9.01	**0.88**

Observed dimension =50



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
- **Finite-sample analysis!** "Sample complexity of interventional causal representation learning". NeurIPS 2024

contact: bvarici@andrew.cmu.edu