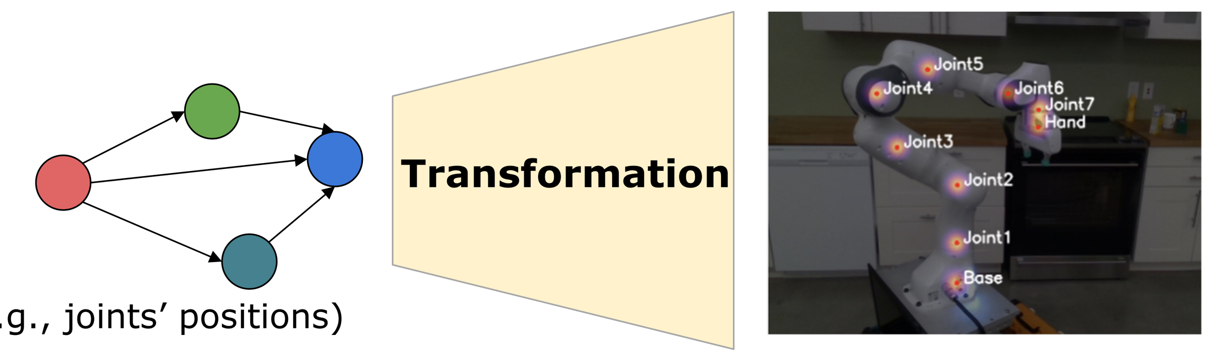
find the paper

Score-based Causal Representation Learning

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

[1]Rensselaer Polytechnic Institute [2]Google Research India

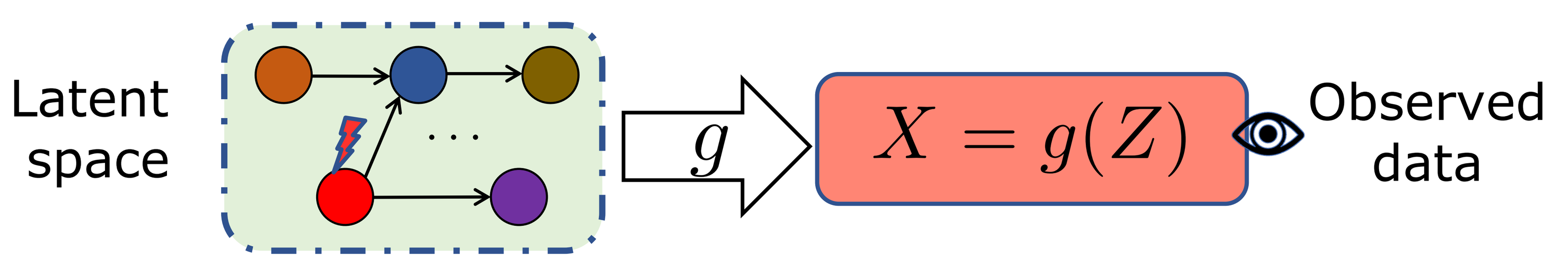
CRL from Interventions



".. learn a representation (partially) exposing the unknown causal structure, e.g., which variables describe the system, and their relations .." Schölkopf et al., 2021

Generic goal: Invert the unknown transformation to recover

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



1. **Identifiability**: Conditions for uniquely recovering Z and $\mathcal{G}_Z$
2. **Achievability**: Provably correct algorithms to recover Z and $\mathcal{G}_Z$

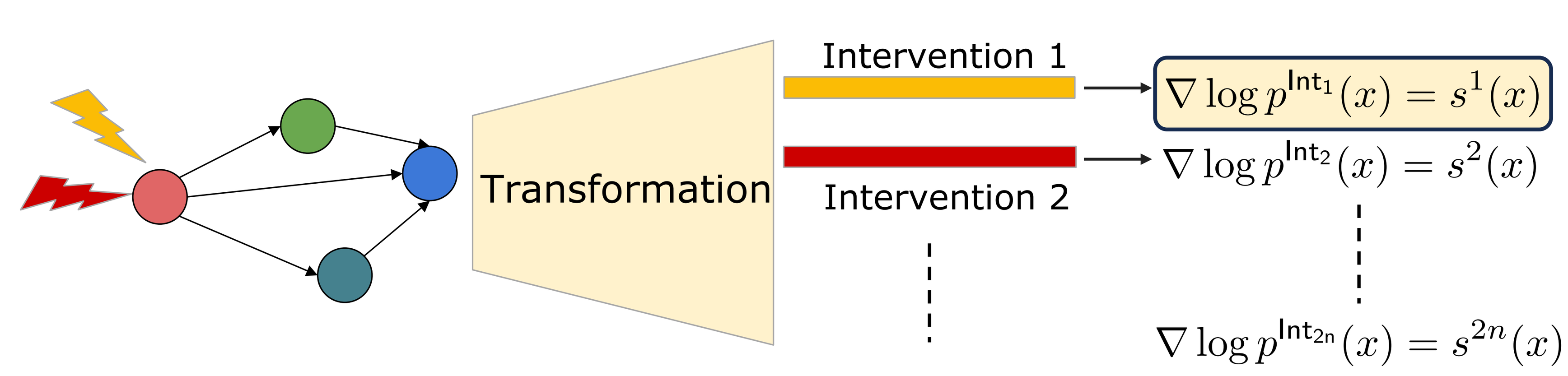
Our contributions

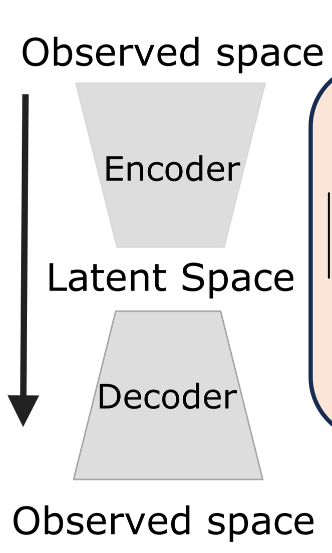
Latent model		Transform		Interv. / node		Main results
Nonparametric	+	Nonparametric	+	2 hard	=	perfect ID
Sufficiently nonlinear	+	Linear	+	1 hard (soft)	=	perfect ID (true DAG + Markov)
Nonparametric	+	Linear	+	1 hard (soft)	=	perfect ID (ID up to ancestors)

provably correct algorithms for all settings

Algorithm Overview

Sufficient Interventional Diversity**:**
2 different hard interventions per node in the latent space





Minimize

$$\Big\| \mathbb{E}\big[|\text{Jac}_{\text{dec}}(x)(s^1(x) - s^2(x))|\big] \ \cdots \ \mathbb{E}\big[|\text{Jac}_{\text{dec}}(x)(s^{2n-1}(x) - s^{2n}(x))|\big] \Big\|_0$$

\+ Reconstruction Loss (ensures injectivity)

Provably correct algorithm for unsupervised learning (small variations for different settings)

Experiments

Non-linear latent model: $Z_i = \sqrt{Z_{\text{pa}(i)}^\top A_{p,i} Z_{\text{pa}(i)}} + N_{p,i}$ n=8 latent variables

Input score differences $(s_X - s_X^m)$: Perfect score oracle or Sliced Score Matching

Non-linear transform: $X = \tanh(T.Z)$

Two hard / node

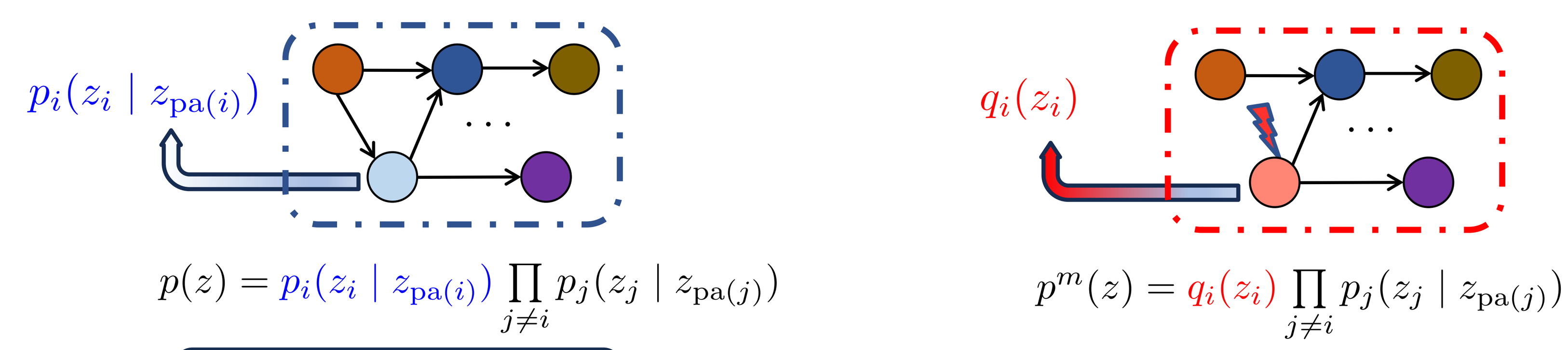
Obs. dim	Norm. Z error (score oracle)	DAG error (SHD) (score oracle)	Norm. Z error (noisy scores)	DAG error (SHD) (noisy scores)
8	0.16	1.56	0.70	11.9
25	0.20	1.55	0.68	10.5
40	0.21	1.14	0.71	11.8

Linear transform: $X = T.Z$

One hard / node

Obs. dim	Norm. Z error (score oracle)	DAG error (SHD) (score oracle)	Norm. Z error (noisy scores)	DAG error (SHD) (noisy scores)
8	0.50	5.4	0.75	10.3
25	0.51	6.0	0.78	8.9
40	0.50	5.3	0.61	11.9

Why score functions?



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

$s(z) \triangleq \nabla_z \log p(z)$ $s^m(z) \triangleq \nabla_z \log p^m(z)$

$$s(z) - s^m(z) = \nabla_z \log p_i(z_i \mid z_{\text{pa}(i)}) - \nabla_z \log q_i(z_i)$$

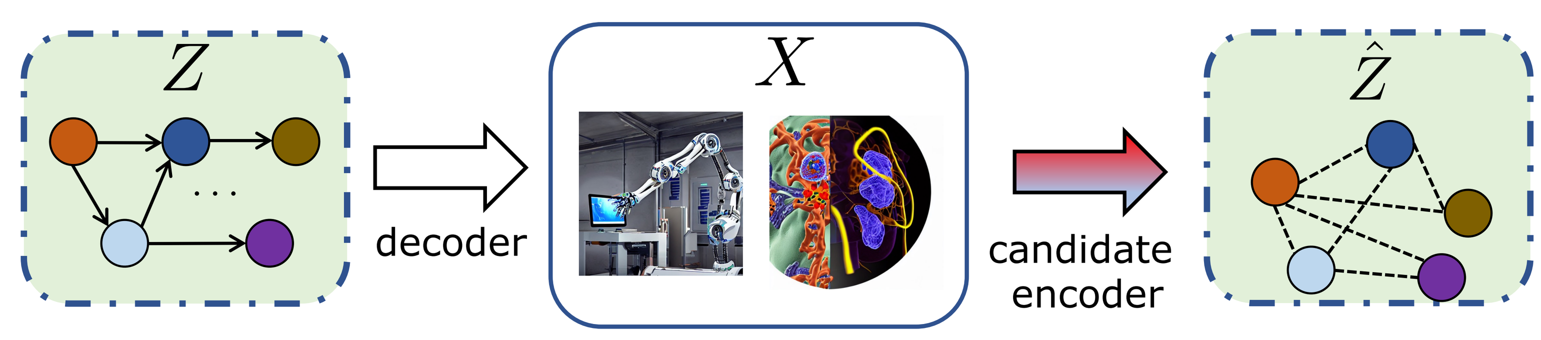
Score functions contain all information about latent DAGs

node i intervened: $s(z) - s^m(z)$ becomes a function of only $z_{\overline{\text{pa}}(i)}$

$$s(z) - s^m(z) = [0 \ \ 0 \ \boxed{\times} \ 0 \ \boxed{\times} \ 0]^\top$$

coordinates of parents of node i

Methodology



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

estimated score differences cannot be sparser than true score differences

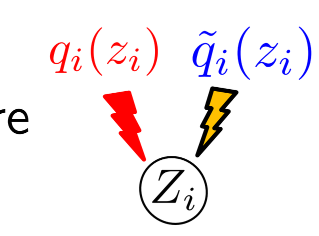
Min. score variations over environment pairs = correct encoder

$$s_{\hat{Z}}(\hat{z}) - s_{\hat{Z}}^m(\hat{z}) = [J_{\text{decoder}}(\hat{z})]^\top (s_X(x) - s_X^m(x))$$

Results

Nonparametric transform

Interventional discrepancy: $\dfrac{\partial}{\partial z_i} \dfrac{q_i(z_i)}{\tilde{q}_i(z_i)} \neq 0$ almost everywhere



Theorem : Observational data and **two hard** interventions/node
Perfect ID

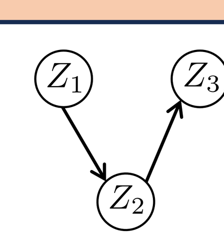
von Kügelgen et al.(2023): Coupled two hard + faithfulness = Perfect ID

Linear transform + nonlinear latents

A1 (nonlinearity): $\text{rank}(\text{im}(s - s^m)) = |\overline{\text{pa}}(i)|$ e.g., 2-layer NN with additive noise

Theorem : Linear transform + **one** intervention/node + **A1**
hard: **Perfect ID ;** **soft**: **Perfect DAG + Markov Property**

Going beyond 'ID ancestors' for soft
(nonlinearity = up to ancestors in Zhang'23)



$\hat{Z}_1 = Z_1$
$\hat{Z}_2 = Z_2$
$\hat{Z}_3 = \text{lin}(Z_2, Z_3)$

Linear transform + any latents

A2 (mild): $\forall j \in \text{pa}(I^m)$, $\frac{[s - s^m]_j}{[s - s^m]_{I^m}} \neq$ constant e.g., weights change in linear model

Theorem : Linear transform + **one** intervention/node + **A2**
hard : **Perfect ID ;** **soft**: **ID up to ancestors**

No parametric restrictions on latents (linear models on Squires'23, Buchholz'23)

reach out at: burakvarici@gmail.com