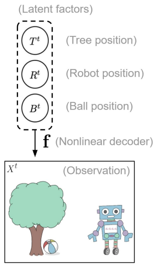
Learning Causal Representations

Apprentissage de représentations causales

Simon Lacoste-Julien

DIRO, Université de Montréal & Mila

Montreal, April 20th 2023

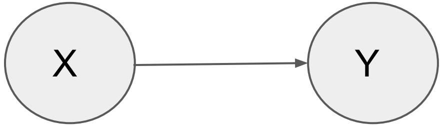


Credits to:

Sébastien Lachapelle

Why causality in machine learning?

- ML researchers are becoming more concerned with **robustness** of their model to **out-of-distribution generalization**
  - One way: use **causal** models
  - Handle **interventions**



- E.g. in reinforcement learning, policy applications, etc. -> need to model the effect of interventions

Learning causal representations

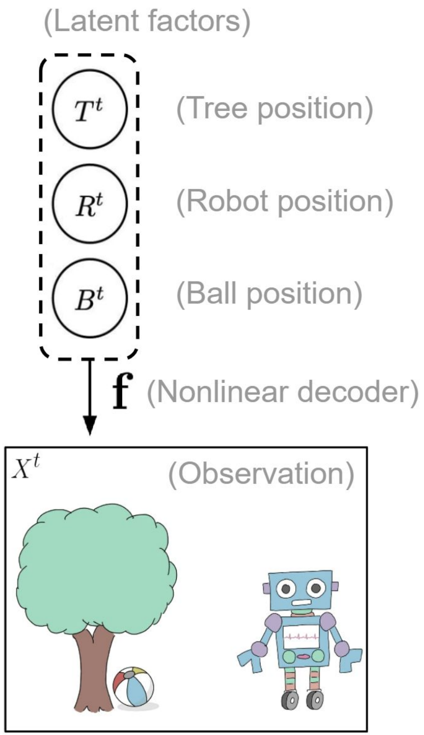
- But what if you don't even know the identity of **causal variables** (high level semantic variables)?
  - -> need to learn them!

Towards Causal Representation Learning

Bernhard Schölkopf †, Francesco Locatello †, Stefan Bauer *, Nan Rosemary Ke *, Nal Kalchbrenner
Anirudh Goyal, Yoshua Bengio

Proceedings of the IEEE 2021

- Motivating example:



Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA

Sébastien Lachapelle[1], Pau Rodríguez López[2], Yash Sharma[3], Katie Everett[4], Rémi Le Priol[1], Alexandre Lacoste[2] and Simon Lacoste-Julien[1]

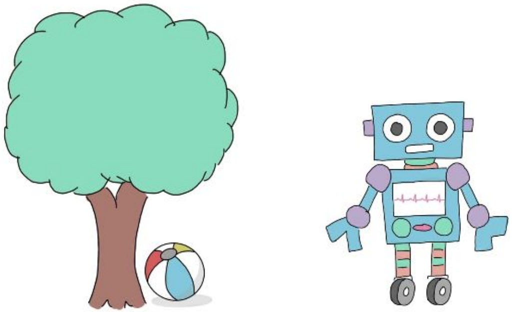
1 Mila, Université de Montréal

2 servicenow

3 Eberhard Karls Universität Tübingen

4 Google Research

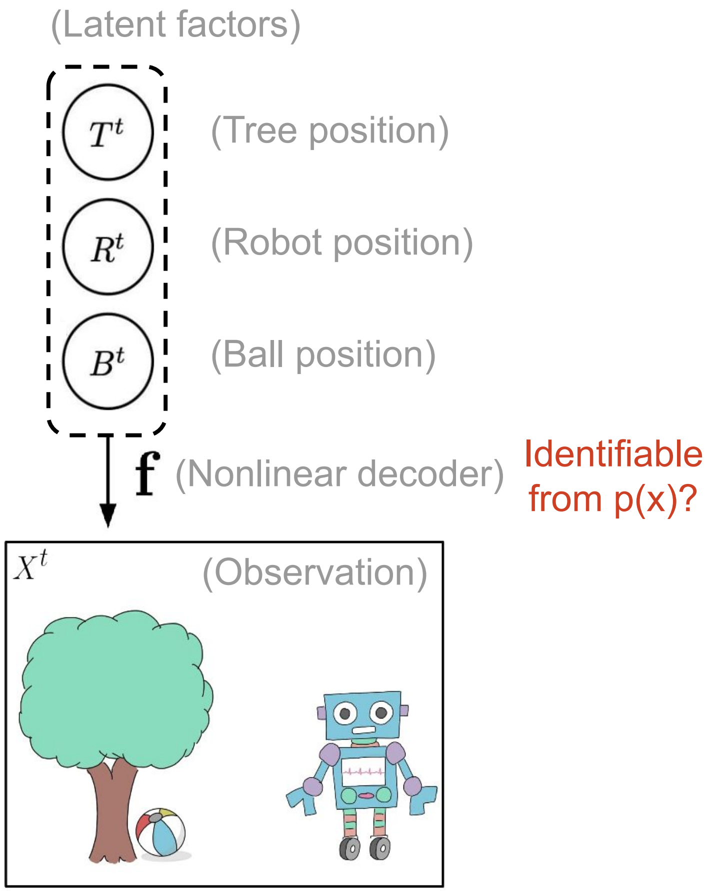
CLeaR 2022



Motivating Example

Disentanglement is the problem of recovering the latent factors without supervision, i.e. from p(x)

Our **identifiability theory** shows how to use this sparsity to disentangle the latent factors via **sparsity regularization**



The general problem of identifiability for generative models

Consider the following simple generative model:

$$Z \sim \mathbb{P}_Z, \; X := \mathbf{f}(Z) \implies \mathbb{P}_X$$

Consider this other model:

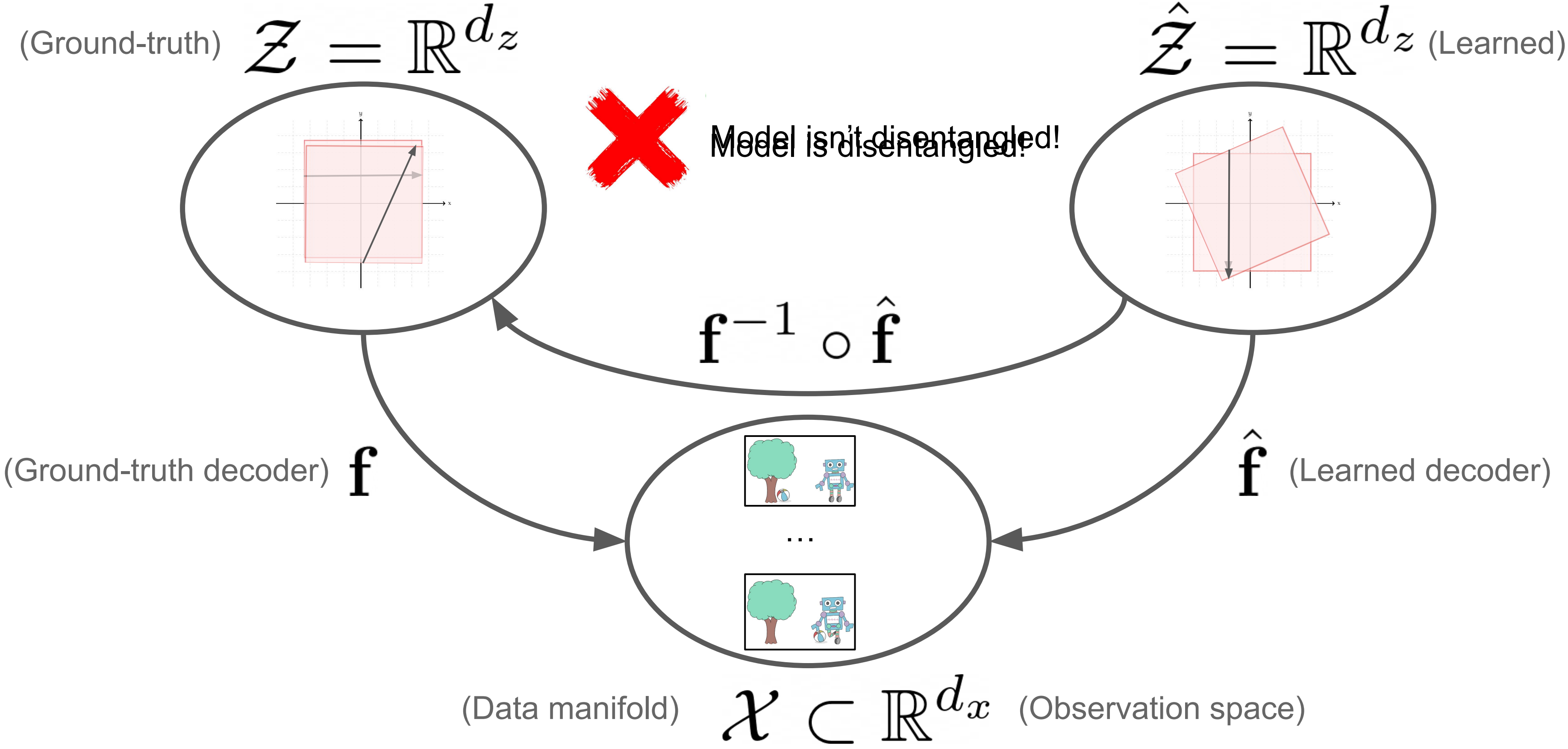
$$\hat{Z} := UZ, \; \hat{X} := \underbrace{\mathbf{f}(U^{-1}}_{\hat{\mathbf{f}}}\hat{Z}) \implies \mathbb{P}_{\hat{X}}$$

=

Both models represent the same distribution over X...

… but their representations can be drastically different

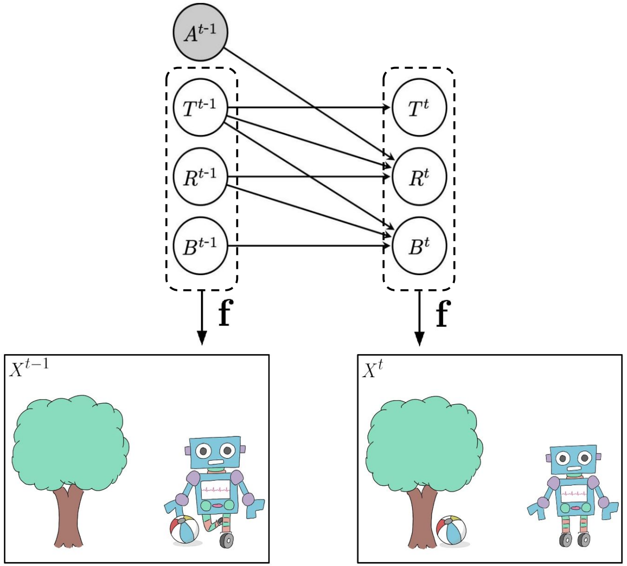
What is disentanglement?



Bird’s-eye view of our approach

- Must learn
  - a decoder $\mathbf{f}$
  - a latent transition model $p(z^t \mid z^{<t}, a^{<t})$ -> parameter $\boldsymbol{\lambda}$
  - Causal graph over the latents/actions G

- Our theory shows how regularizing G to be sparse will help us identify the latent variables, i.e. to have a disentangled representation.



Learnable parameters: $\theta := (\mathbf{f}, \lambda, G)$

Useful to think of

θ = parameter of the ground-truth model

$\hat{\theta}$ or $\tilde{\theta}$ = parameter of the learned model

Our Theorem: Disentanglement via Mechanism Sparsity

(informally)

- If θ and $\hat{\theta}$ model the same distribution on X
- If the ground truth transition is "sufficiently complex"
- If we ensure our predicted graph is **as sparse** as the ground truth graph (this can be done via regularization)
- If the ground truth graph is **sufficiently sparse** (precise in paper)

$$\textit{Then } \theta \textit{ and } \hat{\theta} \textit{ are permutation-equivalent}$$

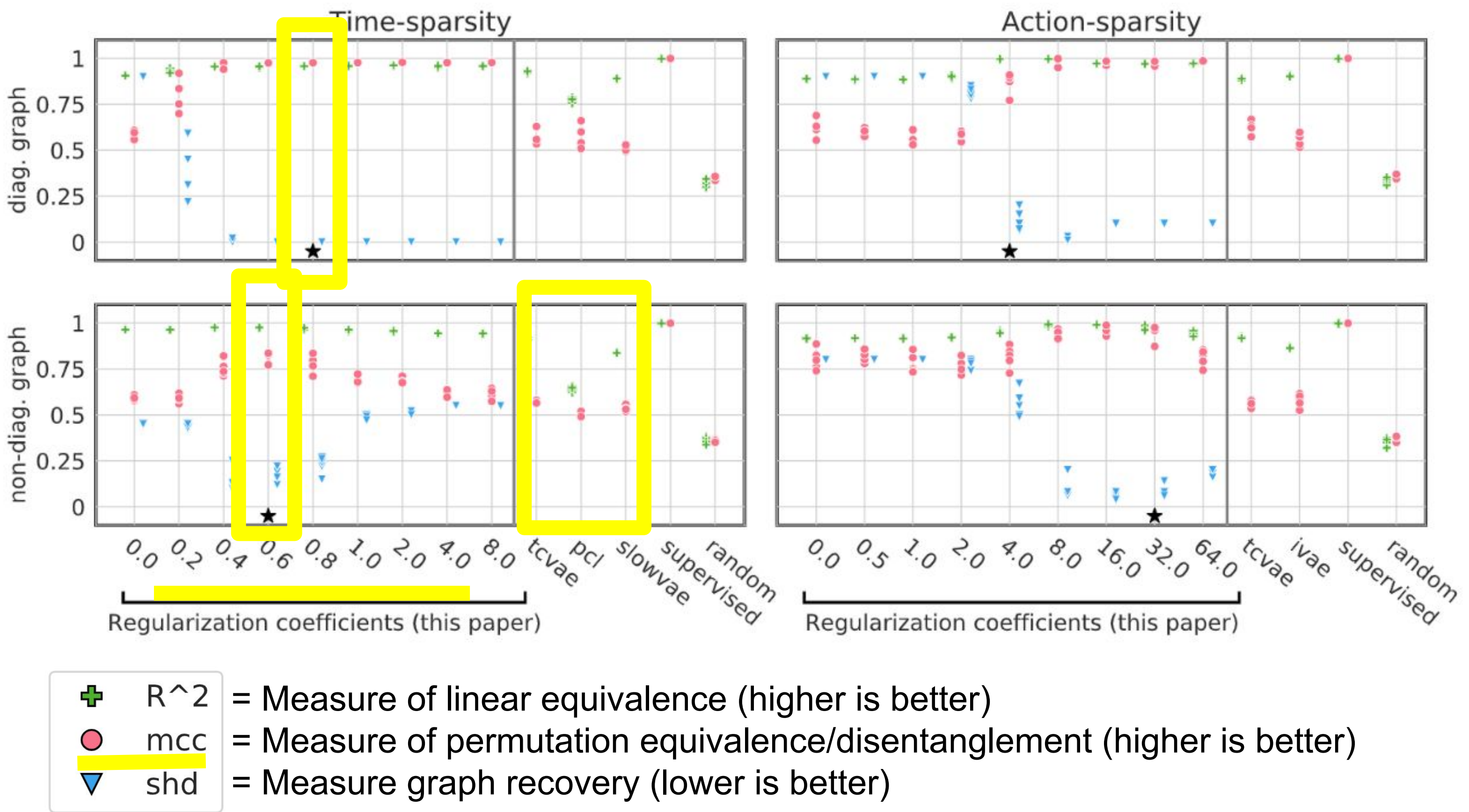
$$\textit{i.e. the model } \hat{\theta} \textit{ is disentangled.}$$

-> Generalized to no constraints on ground truth graph to get **block-permutation equivalence** in paper *"Partial Disentanglement via Mechanism Sparsity"*, [Lachapelle & Lacoste-Julien, arXiv 2022]

Our proposed method

- Model transition model and decoder via neural networks
- Use **binary masks** to encode transition graphs
- Use VAE approach to do approximate maximum likelihood on data
- Use **Gumbel-softmax trick** to learn discrete masks – estimate gradients via reparameterization trick [Jang et al., 2017, Maddison et al., 2017]
  - Can also easily do l0-regularization (sparse graphs) with it

Synthetic Toy Experiments



= Measure of linear equivalence (higher is better)

= Measure of permutation equivalence/disentanglement (higher is better)

= Measure graph recovery (lower is better)

Ongoing work: experiments on Atari games (e.g. pong)

Other work: causal graphical model learning when you know the causal variables

Differentiable Causal Discovery from Interventional Data

(DCDI) NeurIPS 2020

Philippe Brouillard*

Sébastien Lachapelle*

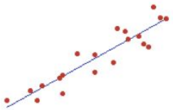
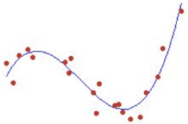
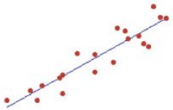
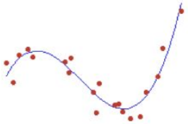
Alexandre Lacoste

Simon Lacoste-Julien

Alexandre Drouin

Structure Learning (from **interventional** data)

Taxonomy of score-based algorithms (non-exhaustive)

			Discrete optim.	Continuous optim.
Observational		Linear	GES [Chickering, 2003]	NOTEARS [Zheng et al., 2018]
		Nonlinear	CAM [Bühlmann et al., 2014]	GraN-DAG [Lachapelle et al., 2020]
Interventional		Linear	GIES [Hauser & Bühlmann, 2012]	-
		Nonlinear	Interventional CAM [Hidden in source code]	**DCDI [Brouillard et al., 2020]**

Conclusion & outlook

- Early steps to learn causal variables from high dimensional data
  - By exploiting sparsity of interactions
- Still need to experiment with more realistic data
- Follow-up work showing the benefit of approach for **multi-task learning** with sparse parameters with sufficient variability across tasks
- Still space for algorithmic improvements, scaling and more theory (finite sample analysis,etc)
- Potential for model-based reinforcement learning

Thank you! – Merci!