# Invariance, Causality and Robustness

Yimeng Chen

Home

April 2020

# Abstract

The "identically and independently distributed" (iid) assumption is referred to as "the big lie in machine learning".

Differences between distribution itself contain valuable information.

# Contents

# Distributional Robustness

For general distributional robustness [3, 5], the aim is to learn

$$\operatorname{argmin}_\theta \sup_{F \in \mathcal{F}} E_F \left( \ell \left( Y, f_\theta(X) \right) \right)$$

for a given set $\mathcal{F}$ of distributions, twice differentiable and convex loss $l$, and prediction $f_\theta(x)$.

The choice of $\mathcal{F}$ is very important as it defined for which environments the model can still be expected to perform well.

# Linear Causal Model

**Causal inference** can be seen as trying to produce models that work well with special distribution classes $\mathcal{F}$.

For **linear** causal model, the solution is exactly the causal parameter [2]

$$\text{argmin}_b \max_{e \in \mathcal{F}} \mathbb{E}\left[|Y^e - X^e b|^2\right] = \text{ causal parameter}$$

where $\mathcal{F} = e$ satisfies 1) $e$ does not act directly on $Y$. 2) $e$ does not change the mechanism between $X^e$ and $Y^e$.

Some forms of interventions are discussed in [5].

# Useful for structure search

This relation opens the door to think about causality in terms of optimizing a certain risk.

This might ease some of the more complicated issues on structure search for causal graphs and structural equation models.

## Problem

How about non-linear case?

# A (related?) work

C. Heinze-Deml et al. [3] splits latent features into "core" and "style", trying to achieve robustness against a set of distributions that are generated by interventions on latent style variables.

They make use of the prior information that some samples come from the same entity (**Counterfactual samples?**). The classification is viewed as anti-causal.

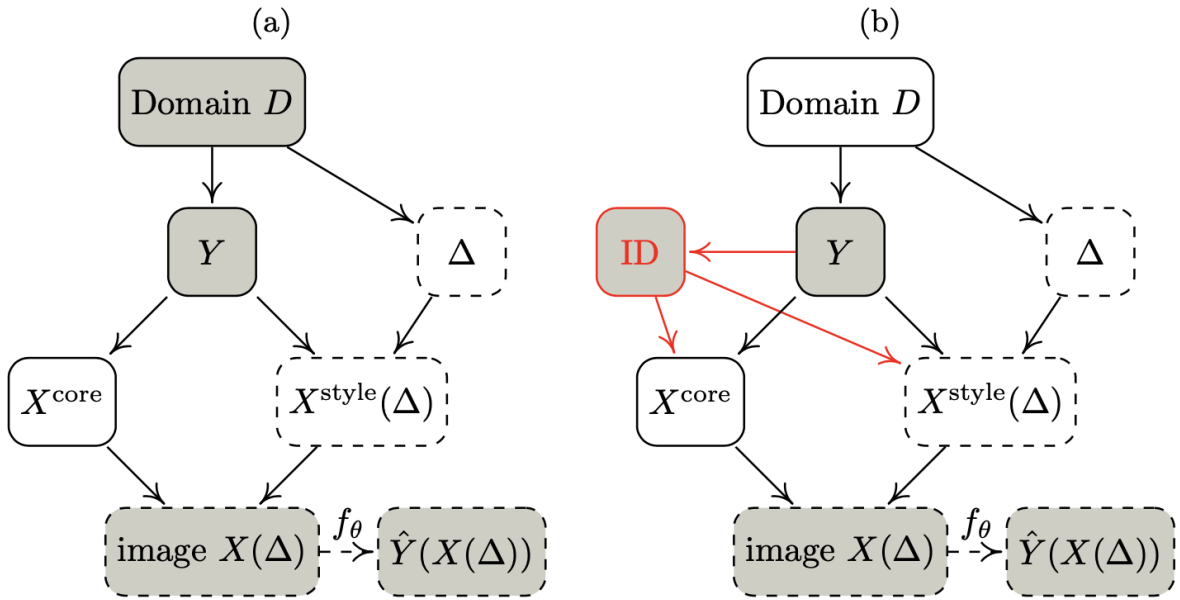


Figure 1: ID: induced. The domain $D$ is latent.

# Section contents

# Invariance Assumption

## Invariance assumption

There exists a subset $S_* \subset 1, ..., p$ of the covariate indices (including the empty set) such that $\mathcal{L}\left(Y^e | X^e_{S_*}\right)$ is the same for all $e \in \mathcal{F}$.

It can be proved that when $\mathcal{F}$ satisfies some causal restriction, the set $S_* = pa(Y)$.

# Invariance and causal discovery

Some work considers the reverse relation Invariance $\Longrightarrow$ causal structures, such as Invariant Causal Prediction (ICP) and Joint Causal Inference (JCI).

# Invariance and transfer learning

Many works are based on this assumption.
It can be viewed as a relaxed version of the usual covariate shift assumption in transfer learning [6], so as to develop the causal transfer learning.
For domain adaptation, S. Magliacane et al. [4] selects invariant features by using JCI.

However this assumption may not suitable in many cases. For example, in CV the input is pixels. This kind of representation is not disentangled.

Problem.
What about NLP?

# Invariance and Data augmentation

When predictions are known to be invariant under some actions (rotations, flipping, etc.), data augmentation can be applied.

When some feature are known to be invariant/variant, counterfactual augmentation can be applied.

# Section contents

# Invariance view of causation

We promote invariance as the main feature of causation.

- The invariance view of causation transcends some of the difficulties of working with causal graphs.
- To find those invariant correlations in machine learning, we need methods which can **disentangle** the observed pixels into latent variables closer to the realm of causation, such as IRM.
- In rare occasions we are truly interested in the entire causal graph governing all the variables in our learning problem. Rather, our focus is often on the causal invariances **improving generalization** across novel distributions of examples.

# Difference between environments

"When shuffling, we destroy information about how the data distribution changes when one varies the data sources or collection specifics. Yet, this information is precisely what tells us whether a property of the data is spurious or stable."

# Algorithms for IRM

Invariant Risk Minimizatio(IRM) is a learning principle to discover unknown invariances from data. The goal is to learn a data representation function $\Phi$, which admits a classifier $w$ simultaneously optimal for all environments $\mathcal{E}$. Also we want $\Phi$ be useful to predict well.
The problem states as

$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} R^e(w\circ\Phi)$$

subject to $\quad w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w}\circ\Phi)$, for all $e\in\mathcal{E}_{\mathrm{tr}}$.
This is a challenging, bi-leveled optimization problem.

# A Practical Version of IRM

A practical version (IRMv1):

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} R^e(\Phi) + \lambda \cdot \left\|\nabla_{w|w=1.0} R^e(w\cdot\Phi)\right\|^2$$

The classifier $w$ is assumed to be linear. The derivation from IRM to IRMv1 takes several steps.

# Invariance, causality and generalization

One can show that a predictor $v : X \to Y$ is invariant across $\mathcal{E}_{all}$ if and only if it attains minimal risk, and if and only if it uses only the direct causal parents of Y to predict.
The generalization ability of IRM is derived under the following assumptions:

- Linear relationship.
- Some degree of diversity across training environments.

# Experiment: Colored MNIST

Images are labeled 0 for digits 0-4 and 1 for 5-9.
Each image is colored either red or green in a way that correlates strongly (but spuriously) with the class label.

| **Algorithm** | **Acc. train envs.** | **Acc. test env.** |
|---|---|---|
| ERM | $87.4 \pm 0.2$ | $17.1 \pm 0.6$ |
| **IRM (ours)** | $70.8 \pm 0.9$ | $\mathbf{66.9 \pm 2.5}$ |
| Random guessing (hypothetical) | 50 | 50 |
| Optimal invariant model (hypothetical) | 75 | 75 |
| ERM, grayscale model (oracle) | $73.5 \pm 0.2$ | $73.0 \pm 0.4$ |

Figure 2: Accuracy of different algorithms on the Colored MNIST synthetic task.

# Causal or Anti-causal?

Some researchers hold the idea that classification task is typically anti-causal (Bernhard Scholkopf et al., and [3]).

However M. Arjovsky, Leon Bottou et al. [1] believes that **most supervised learning problems**, such as image classification, are **causal**:

If the **annotation process** is close to deterministic and shared across environments, predicting annotations is a causal problem. It models human cognition process.

# Causal or Anti-causal?

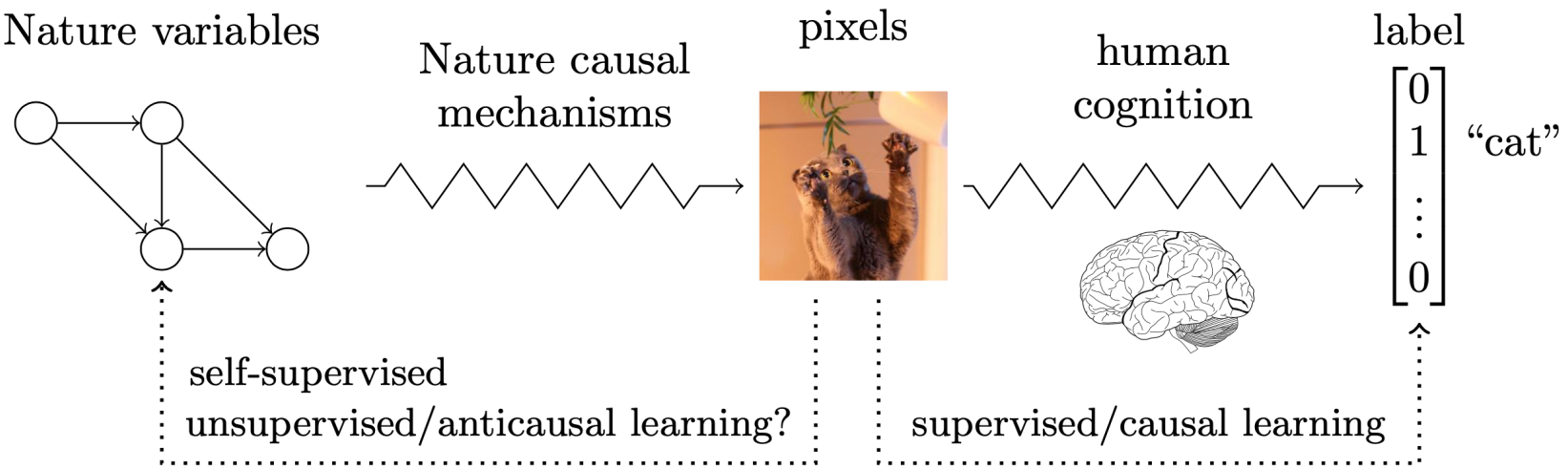


Figure 3: All learning problems use empirical observations, here referred to as "pixels".

# Future work

1. "nonlinear general position" assumption and prove that it holds almost everywhere.
2. What problems allow the discovery of invariances from **few** environments?
3. Perhaps we could think of reinforcement learning episodes as different environments, so we can learn **robust policies** that leverage the invariant part of behaviour leading to reward.
4. It may be possible to formalize IRM in terms of invariance and equivariance concepts from **group theory**.

# Math tools for Invariance

- Topology (may be more suitable for CV? NLP is more algebraic.)
- Group theory, which studies the invariance under group action.
- Algebraic geometry?

Thanks for your attention!

# Related Works

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019.

[2] Peter Bühlmann. Invariance, causality and robustness, 2018.

[3] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness, 2017.

[4] Sara Magliacane. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems 31*. 2018.

[5] N. Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, 2018.

[6] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *J. Mach. Learn. Res.*, 2018.