Towards Causal Representation Learning

Based on paper https://arxiv.org/pdf/2102.11107.pdf

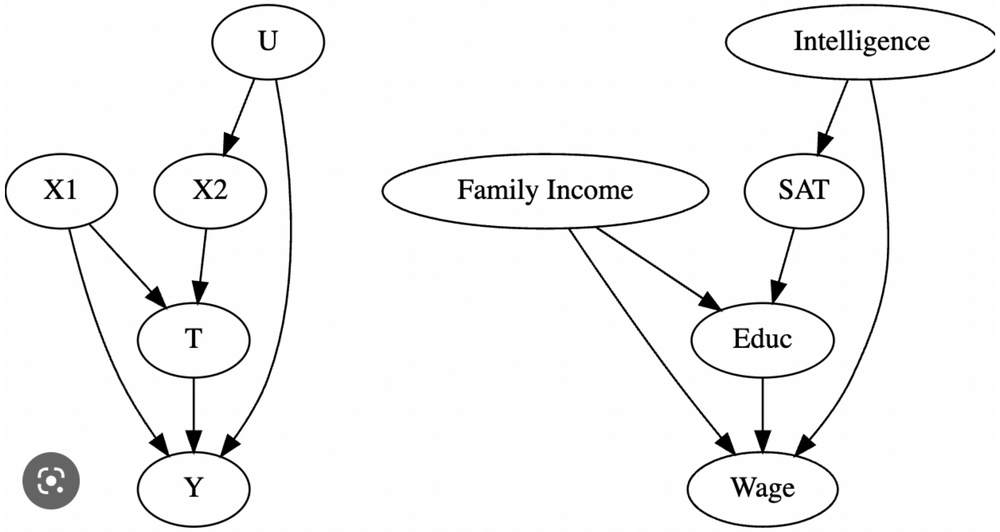
Presented by

Chang Ma

Contents

- Why do we need causal representation learning
  - Out-of-distribution generalization
- Fundamentals of causal representation learning
- Causal Discovery and Machine Learning
- How to learn

Levels of Causal Reasoning Ability



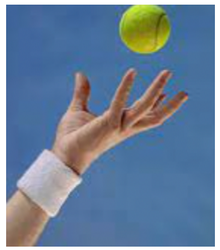
Layer	Activity	Semantics	Example
(1) Associational $p(y \mid x)$	Seeing 👀	How would seeing x change my belief in Y?	Is wage related to intelligence
(2) Interventional $p(y \mid \mathrm{do}(x), z)$	Doing 💪	What happens to Y if I do x?	What if he is more clever?
(3) Counterfactual $p(y_{x'} \mid x, y)$	Imagining 🤔	Was it x that caused Y?	Was it intelligence that leads to low income?

Causal for OOD Generalization

- Current machine learning = large scale pattern recognition on suitably collected i.i.d. data
    - Often fail on robustness
    - Hard for domain generalization
    - Fail to Learning Reusable Mechanisms

Causal for OOD Generalization

- 1. Defined based on interventions rather than correlation
    - Conditional probability from data: (“seeing people with open umbrellas suggests that it is raining”)
    - Outcome of an active intervention (“closing umbrellas does not stop the rain”)
    - Avoid false generation
- 2. Causal relations as reasoning chains, provide predictions for situations far from seen distributions
    - Because how a+b is defined, unseen calculations can be implemented
- 3. Compositionality improves systematic generalization
    - Involves abstractive reasoning
    - Dynamically recombine existing concepts



A thrown ball drops because of gravity.

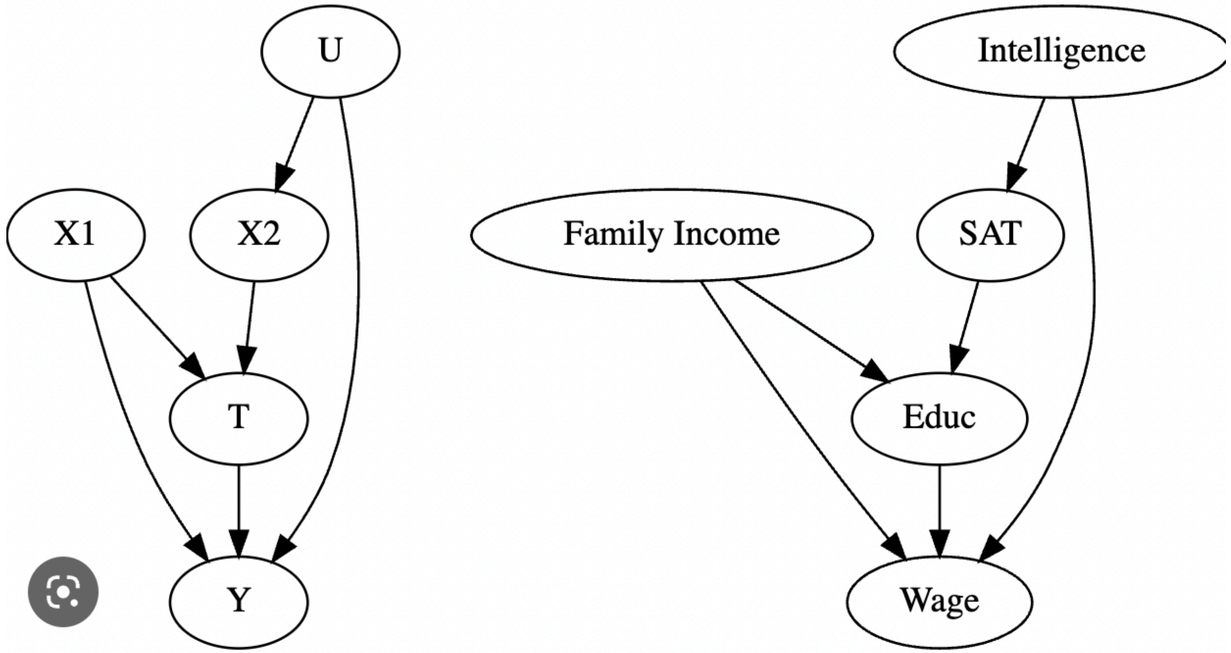
Concept: dropped objects VS changes in pixels

Causality for Generalization

Model	Predict in i.i.d. setting	Predict under distr. shift/intervention	Answer counter-factual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

- Mechanistic: $\frac{d\mathbf{x}}{dt} = f(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^d,$ dynamic systems
- Structural causal VS Causal graphical

Structural causal Model
VS Causal graphical Model



Structural causal models

$$X_i := f_i(\mathbf{PA}_i, U_i), \qquad (i = 1, \ldots, n)$$

Causal graphical model (disentangled factorization)

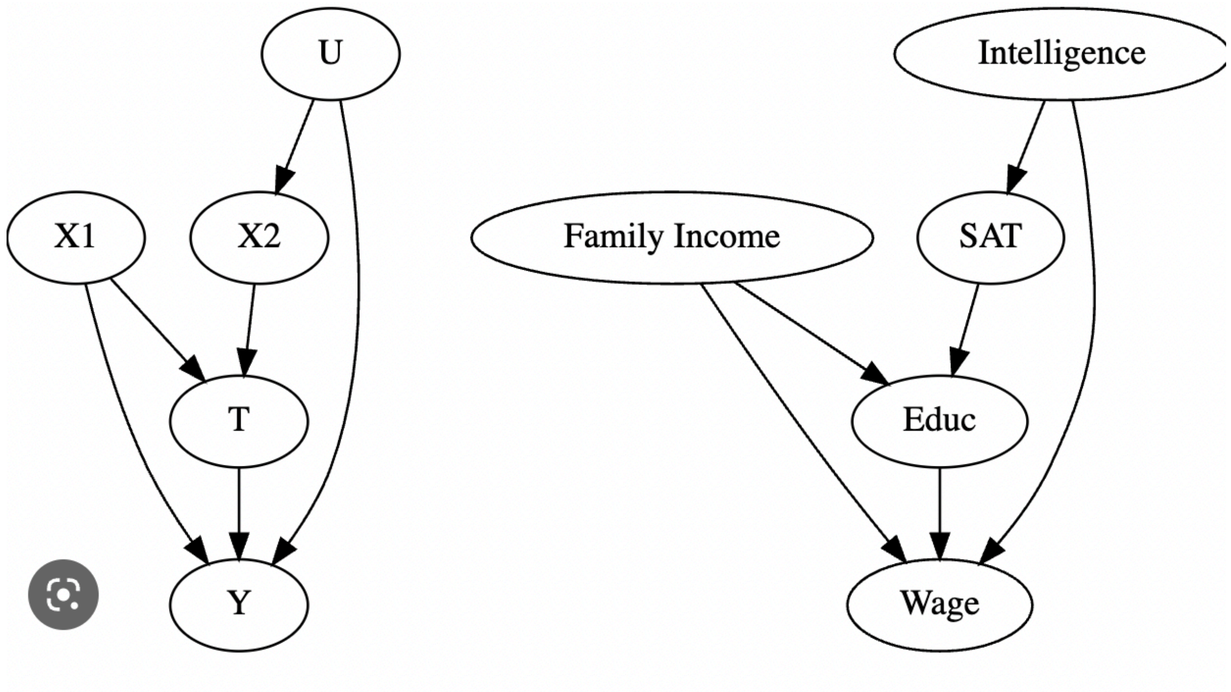
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathbf{PA}_i).$$

(VS entangled factorization)

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{i+1}, \ldots, X_n),$$

Cannot solve counterfactuals

Counterfactual based on SCM



Structural causal models

$$X_i := f_i(\mathbf{PA}_i, U_i), \qquad (i = 1, \ldots, n)$$

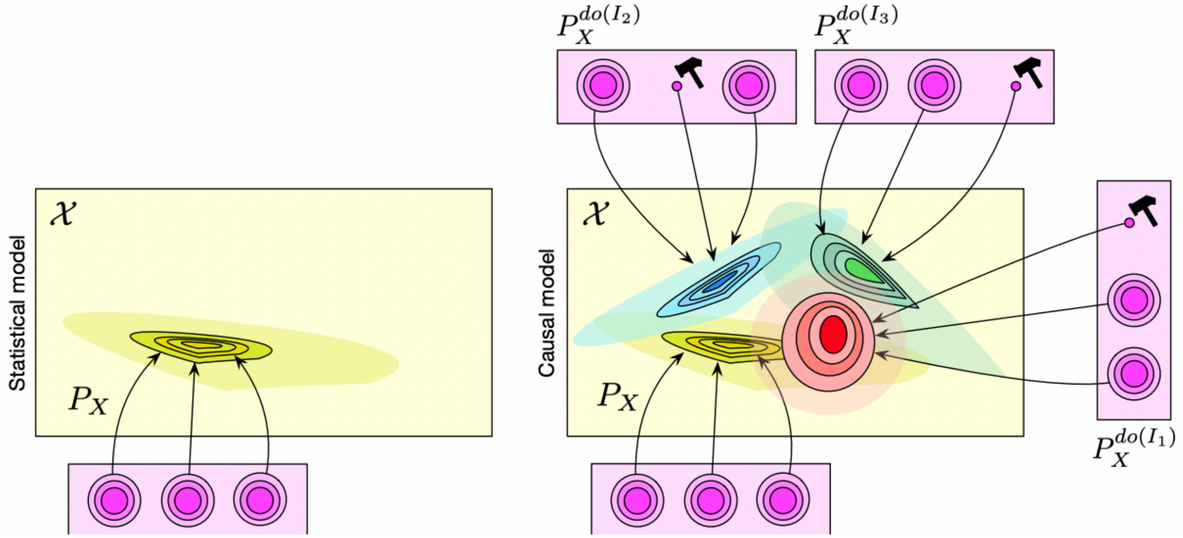
- Abductive reasoning: based on observable X1,X2, infer other underlying variable determining these observables P(U|X1,X2). Then build prediction model based on sampled underlying variables, u~P(U|X2,Y), F = P(Y|U,X1,X2) with MLE
- Action: Assign intervention values to X1 variable. New F’
- Prediction with P(U|X1,X2) and F’

Intervention based on SCM

Structural causal models

$$X_i := f_i(\mathbf{PA}_i, U_i), \qquad (i = 1, \ldots, n)$$

Independent Causal Mechanisms (ICM) Principle. *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.*

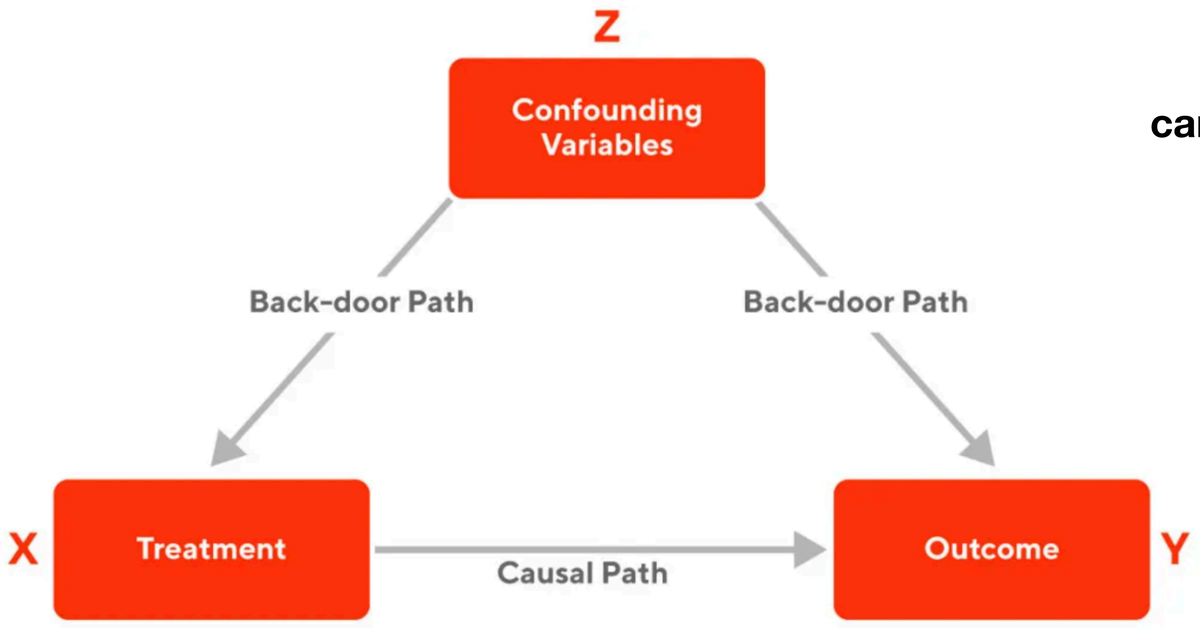


a causal model represents a set of distributions

Intervention can be applied independently to disentangled variables

Confounder Effect and Backdoor Adjustment

Structural causal models



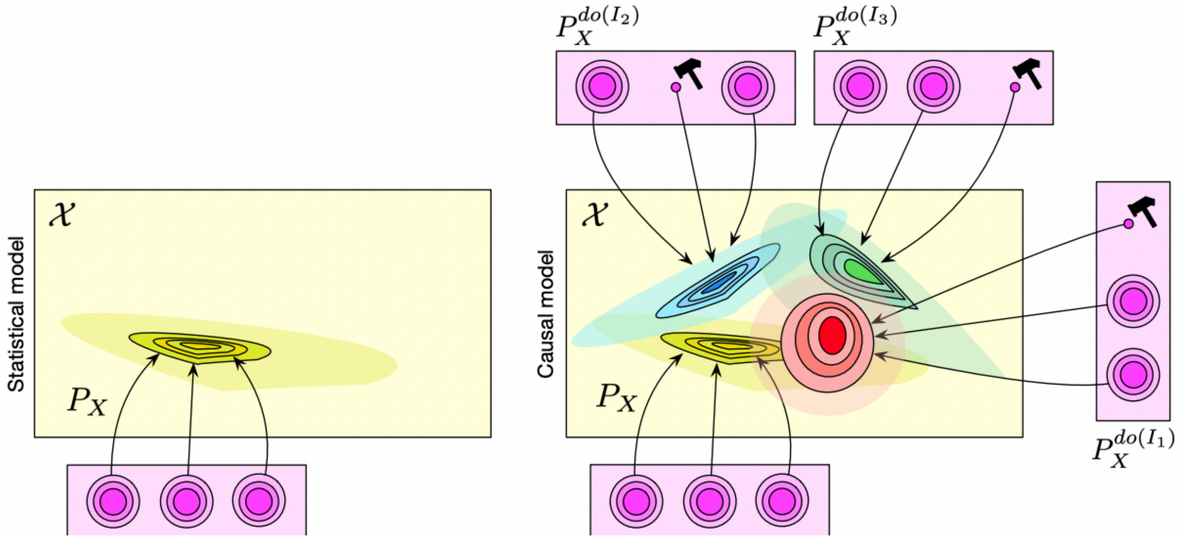
Intervention on confounders Z,
can measure the causal relationship between X & Y

Confounder Effect and Backdoor Adjustment

Structural causal models

Independent Intervention on different variables
Generate a set of distributions.

True causality would find high correlation in all conditional distributions.

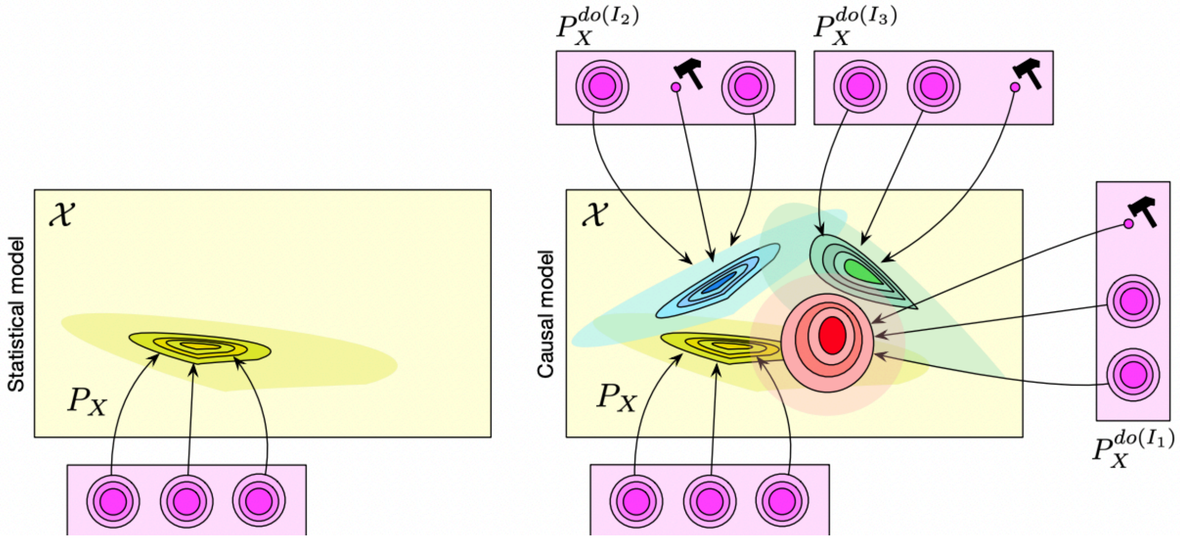


Confounder Effect and Backdoor Adjustment

Structural causal models

Independent Intervention on different unseen variables
Generate a set of distributions.

True causality would find high correlation in all conditional distributions.

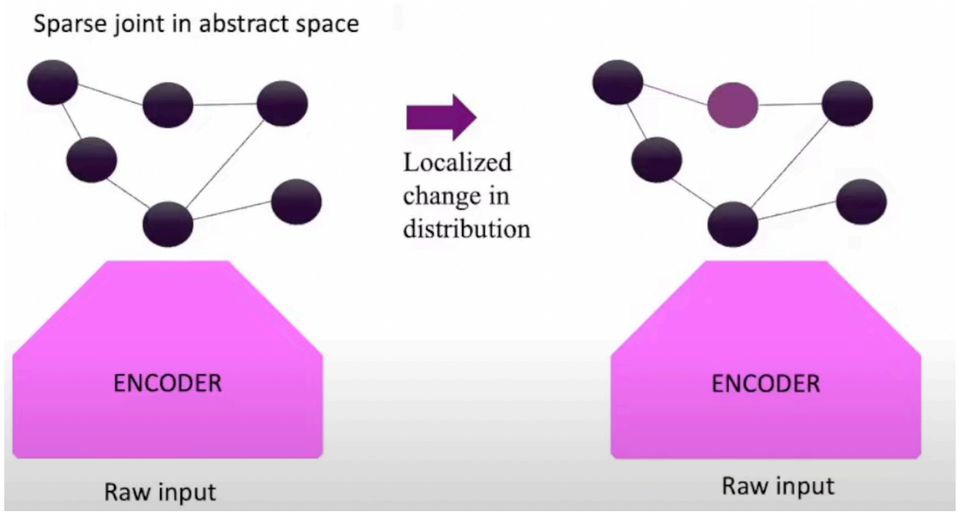


Independent Causal Mechanisms (ICM) Principle.
The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.

P(A,B) = P(A)P(B|A), When A ➡️ B, P(B|A) remains stable, and this phenomenon does not hold when it is reversed.

Distribution changes due to localized intervention

Sparse Mechanism Shift (SMS). *Small distribution changes tend to manifest themselves in a sparse or local way in the causal/disentangled factorization (4), i.e., they should usually not affect all factors simultaneously.*



- Change maybe drastic in input space
- Yet minimal in semantic space of causal variables
- Hypothesis to replace i.i.d: generalization = consequence of an intervention on few causes/mechanisms

Causal view for ML

- Semi-supervised learning
- $P(X, Y) = P(X)P(Y|X).$
- Based on ICM principles, P(X) should contain no information about P(Y |X), which implies that SSL should be futile, in as far as it is using additional information about P(X) (from unlabelled data) to improve our estimate of P(Y |X = x).

Causal view for ML

- Adversarial vulnerability
  - attacks clearly constitute violations of the i.i.d. assumption that underlies statistical machine learning
  - Traditionally solves this with lp ball constraints
  - Hard to sample all attacks possible
  - Based on SMS principle, need only guard against interference in a sparse space with only a few causes (e.g. defense against attacks in representation space, if better, disentangled space.)

Causal view for ML

- OOD generalization
  - Spurious correlations (confounder effect)
  - Group distributional robust optimization

with interventions

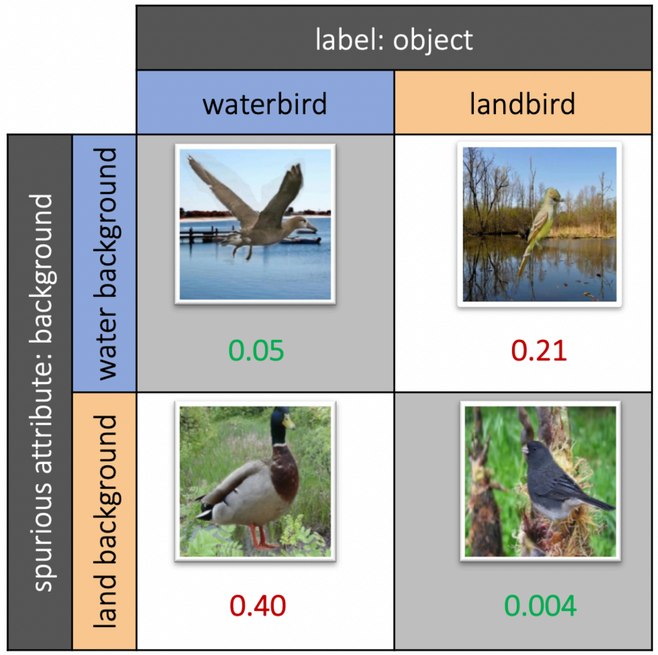
Using causal language, one could restrict $P^\dagger(X,Y)$ to be the result of a certain set of interventions, i.e., $P^\dagger(X,Y) \in \mathbb{P}_\mathcal{G}$ where $\mathbb{P}_\mathcal{G}$ is a set of interventional distributions over a causal graph $\mathcal{G}$. The worst case out-of-distribution risk then becomes

$$R^{OOD}_{\mathbb{P}_\mathcal{G}}(g) = \max_{P^\dagger \in \mathbb{P}_\mathcal{G}} \mathbb{E}_{P^\dagger(X,Y)} \left[\text{loss}(Y, g(X))\right]. \qquad (17)$$

To learn a robust predictor, we should have available a subset of environment distributions $\mathcal{E} \subset \mathbb{P}_\mathcal{G}$ and solve

$$g^\star = \operatorname*{argmin}_{g \in \mathcal{H}} \max_{P^\dagger \in \mathcal{E}} \hat{\mathbb{E}}_{P^\dagger(X,Y)} \left[\text{loss}(Y, g(X))\right]. \qquad (18)$$

-



worst-group error: 0.40

Future work

- 1. Disentangled representations
- 2. Extracting causal graph (from unstructured data)
- 3. Causal in science