

Mining the Data Heterogeneity for Out-of-Distribution Generalization

Jiashuo Liu

TrustWorthy-AI Group, CST, Tsinghua University

2022.08.31



① Causality & Invariance & Heterogeneity

② Invariant Learning Problem under Latent Heterogeneity

③ Distributional Invariance Property

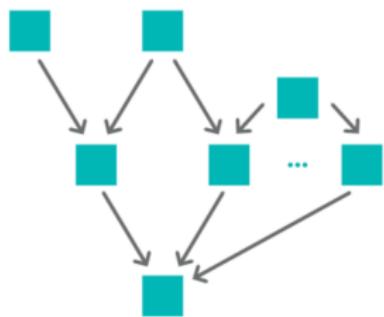
① Causality & Invariance & Heterogeneity

② Invariant Learning Problem under Latent Heterogeneity

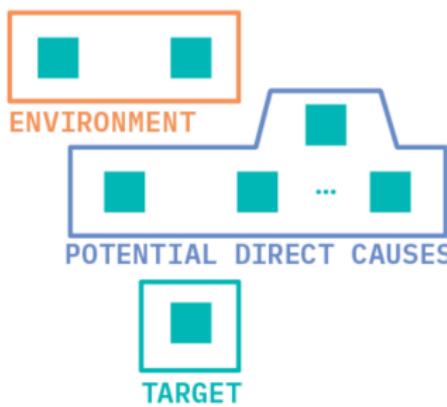
③ Distributional Invariance Property

Causality & Invariance

Full Causal Graph



Fewer Assumptions



1

¹Causality for Machine Learning. Cloudera's Fast Forward Labs

Invariance Property

There are several versions of the **Invariance Assumption**.

Assumption (Invariance Assumption²)

There exists random variable $\Phi(X)$ such that for all $e_1, e_2 \in \text{supp}(\mathcal{E})$, we have

$$P^{e_1}(Y|\Phi(X)) = P^{e_2}(Y|\Phi(X)) \quad (1)$$

- *This assumption is equivalent to $Y \perp \mathcal{E} | \Phi(X)$, indicating that the relationship between $\Phi(X)$ and Y remains invariant across environments, which is also referred to as causal relationship.*

Assumption (Invariance Assumption³)

There exists random variable $\Phi(X)$ such that for all $e_1, e_2 \in \text{supp}(\mathcal{E})$, we have

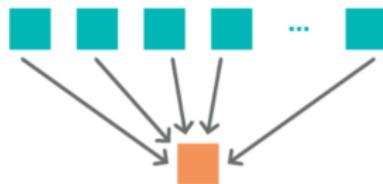
$$\mathbb{E}^{e_1}[Y|\Phi(X)] = \mathbb{E}^{e_2}[Y|\Phi(X)] \quad (2)$$

²Koyama, Masanori, and Shoichiro Yamaguchi. "Out-of-distribution generalization with maximal invariant predictor." (2020).

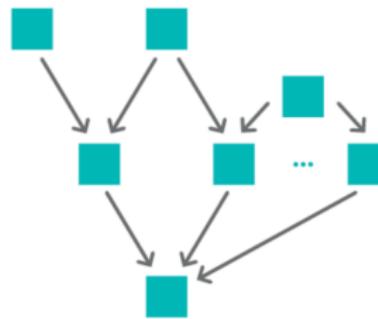
³Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint arXiv:1907.02893.

Advantage on the Out-of-Distribution Generalization Problem

Machine Learning



Casuality



For prediction problem, the invariance property is enough (we only care about $\text{Pa}(Y)$). \Rightarrow Do not need the whole causal graph.

- $\Phi^*(X) = \arg \max_{\Phi: Y \perp \mathcal{E} | \Phi} \mathbb{I}(Y; \Phi(X))$ is referred to as **(Maximal) Invariant Predictors**.
- Under some assumptions, $\mathbb{E}[Y|\Phi^*(X)]$ can achieve OOD optimality⁴.

⁴Koyama, Masanori, and Shoichiro Yamaguchi. "Out-of-distribution generalization with maximal invariant predictor." (2020).

Advantage on the Out-of-Distribution Generalization Problem

- **Out-of-Distribution Generalization Problem(OOD Problem)** is proposed in order to guarantee the generalization ability under distributional shifts, which can be formalized as:

$$\theta_{OOD} = \arg \min_{\theta} \max_{e \in \text{supp}(\mathcal{E})} \mathcal{L}^e(\theta; X, Y) \quad (3)$$

where

- \mathcal{E} is the random variable on indices of all possible environments, and for each environment $e \in \text{supp}(\mathcal{E})$, the data distribution is denoted as $P^e(X, Y)$.
- The data distribution $P^e(X, Y)$ can be quite different among environments in $\text{supp}(\mathcal{E})$.
- $\mathcal{L}^e(\theta; X, Y)$ denotes the risk of predictor θ on environment e , whose formulation is given by:

$$\mathcal{L}^e(\theta; X, Y) = \mathbb{E}_{X, Y \sim P^e} [\ell(\theta; X, Y)] \quad (4)$$

Invariance Property

Invariance to What? \Rightarrow Some limitations in practice.

Env1



Env2



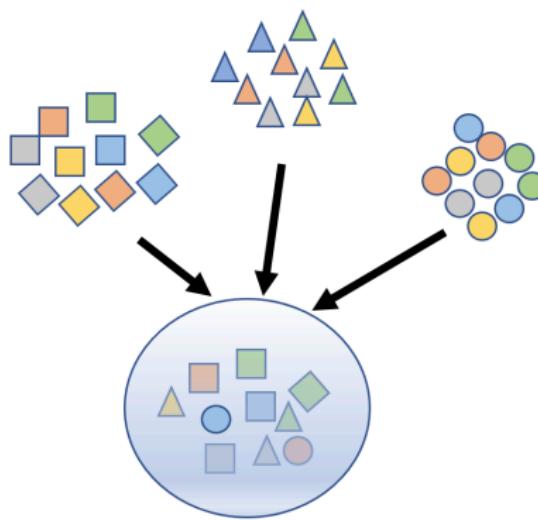
Env3



- When environment set \mathcal{E} contains *Env1* and *Env2*: grass is invariant.
- When environment set \mathcal{E} contains *Env1* and *Env3*: grass is variant.

Limitation 1: No environment labels

Modern datasets are frequently assembled by merging data from multiple sources **without explicit source labels**, which means there are not multiple environments but only one pooled dataset.



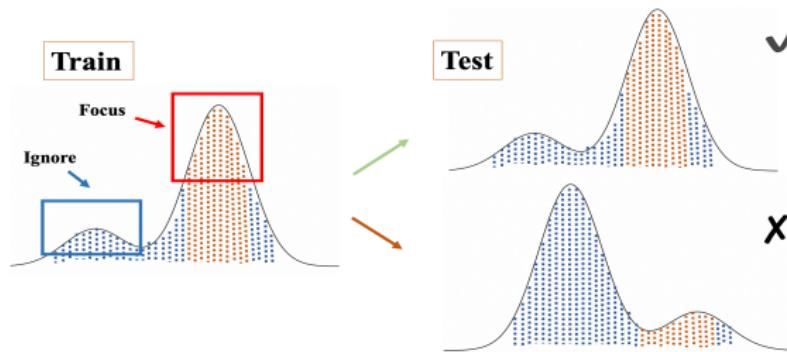
Limitation 2: Quality of environments

- **Heterogeneous Enough?**
 - whether environments are heterogeneous to reveal the variant relationships
 - for example, all environments are the same \Rightarrow useless
- **Homogeneous Enough?**
 - whether the invariance holds among the environments
 - for example, some environments are polluted, and only random noises Φ satisfies $Y \perp \mathcal{E} | \Phi \Rightarrow$ useless

Heterogeneity

Data are collected from multiple sources, which induces latent heterogeneity.

- ERM excessively focuses on the majority and ignores the minor components in data.
- Overall Good = **Majority Perfect** + **Minority Bad**
- Majority and Minority can change across different data sources/environments.
- Latent Heterogeneity renders ERM break down under distributional shifts.



Insights: We should leverage the latent heterogeneity in data and develop more rational risk minimization approach to achieve **Majority Good** and **Minority Good**, resulting in our **Invariant Learning Problem under Latent Heterogeneity**.

Leverage the Heterogeneity to Learn Invariance

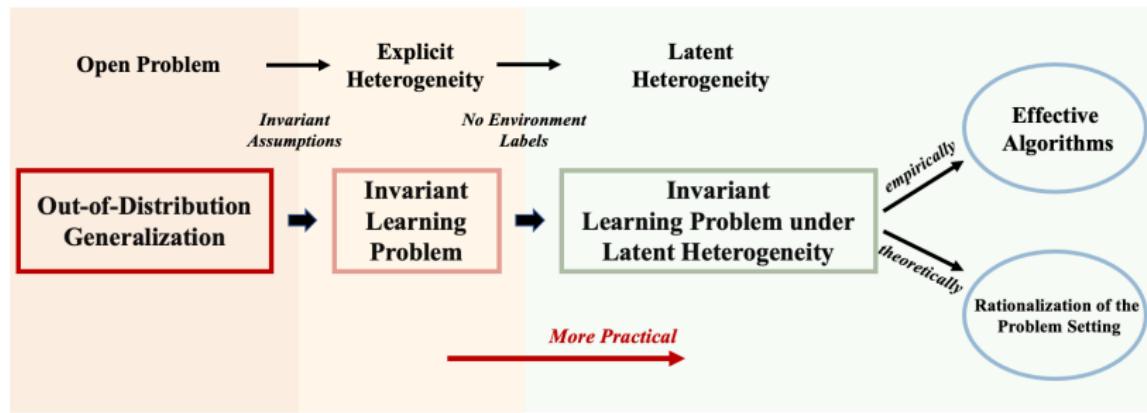
Another compelling but untested option is to try combining IRM with some sort of clustering to segment a single dataset into environments.^[37] The question would be how to cluster in such a way that meaningful and diverse environments are defined. Since existing clustering approaches are purely correlative, and - as such - vulnerable to spurious correlations, this could prove challenging.

Studying the impact of environment selection, and how to create or curate datasets with multiple environments would be a valuable contribution to making invariance-based methods more widely applicable. (The authors of An Empirical Study of Invariant Risk Minimization reach the same conclusion.)

5

⁵Causality for Machine Learning. Cloudera's Fast Forward Labs

An Overview



① Causality & Invariance & Heterogeneity

② Invariant Learning Problem under Latent Heterogeneity

③ Distributional Invariance Property

Invariant Learning Problem under Latent Heterogeneity

Assumption (Heterogeneity Assumption)

For random variable pair (X, Φ^*) and Φ^* satisfying the Invariance Assumption, using functional representation lemma⁶, there exists random variable Ψ^* such that $X = X(\Phi^*, \Psi^*)$, then we assume $P^e(Y|\Psi^*)$ can arbitrary change across environments $e \in \text{supp}(\mathcal{E})$.

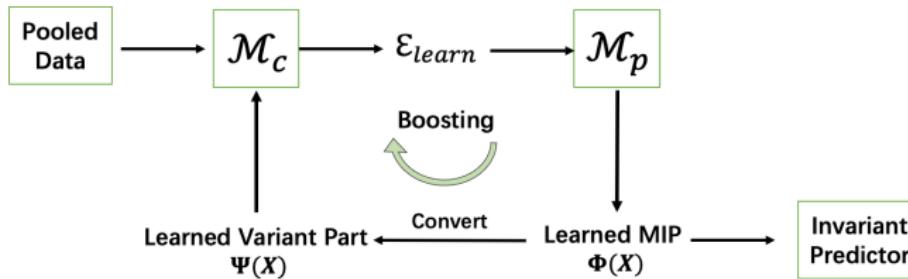
Problem (Invariant Learning Problem under Latent Heterogeneity)

Given heterogeneous dataset $D = \{D^e\}_{e \in \text{supp}(\mathcal{E}_{\text{latent}})}$ without environment labels, the task is to generate environments $\mathcal{E}_{\text{learn}}$ and learn invariant model under learned $\mathcal{E}_{\text{learn}}$ with good OOD performance.

⁶El Gamal, A. and Kim, Y.-H. Network information theory. Network Information Theory, 12 2011.

Empirical Algorithm 1: Heterogeneous Risk Minimization⁷

- This work temporarily focuses on a simple but general setting, where $X = [\Phi^*, \Psi^*]^T$ at the raw feature level.
- The HRM framework contains two modules, named **Heterogeneity Identification** module \mathcal{M}_c and **Invariant Prediction** module \mathcal{M}_p .



- The two modules can **mutually promote** each other, meaning that the invariant prediction and the quality of E_{learn} can both get better and better.
- We adopt feature selection to accomplish the conversion from $\Phi(X)$ to $\Psi(X)$.
- Under our raw feature setting, we simply let $\Phi(X) = M \odot X$ and $\Psi(X) = (1 - M) \odot X$.

⁷ Jiashuo Liu, Zheyuan Hu, Peng Cui *et al.* Heterogeneous Risk Minimization. *In ICML 2021.*

The Heterogeneity Identification Module \mathcal{M}_c

Recall that for \mathcal{M}_c ,

$$\Psi(X) \rightarrow \mathcal{M}_c \rightarrow \mathcal{E}_{learn}$$

we implement it with a convex clustering method. Different from other clustering methods, we cluster the data according to the **relationship** between $\Psi(X)$ and Y .

- Assume the j -th cluster centre $P_{\Theta_j}(Y|\Psi)$ parameterized by Θ_j to be a Gaussian around $f_{\Theta_j}(\Psi)$ as $\mathcal{N}(f_{\Theta_j}(\Psi), \sigma^2)$:

$$h_j(\Psi, Y) = P_{\Theta_j}(Y|\Psi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y - f_{\Theta_j}(\Psi))^2}{2\sigma^2}\right) \quad (5)$$

- The empirical data distribution is $\hat{P}_N = \frac{1}{N} \sum_{i=1}^N \delta_i(\Psi, Y)$
- The target is to find a distribution in $\mathcal{Q} = \{Q | Q = \sum_{j \in [K]} q_j h_j(\Psi, Y), \mathbf{q} \in \Delta_K\}$ to fit the empirical distribution best.
- The objective function of our heterogeneous clustering is:

$$\min_{Q \in \mathcal{Q}} D_{KL}(\hat{P}_N \| Q) \quad (6)$$

The Invariant Prediction Module \mathcal{M}_p

Recall that for \mathcal{M}_p ,

$$\mathcal{E}_{\text{learn}} \rightarrow \mathcal{M}_p \rightarrow \Phi(X) = M \odot X$$

The algorithm involves two parts, invariant prediction and feature selection.

- For invariant prediction, we adopt the regularizer⁸ as:

$$\mathcal{L}_p(M \odot X, Y; \theta) = \mathbb{E}_{\mathcal{E}_{\text{tr}}}[\mathcal{L}^e] + \lambda \text{trace}(\text{Var}_{\mathcal{E}_{\text{tr}}}(\nabla_{\theta} \mathcal{L}^e)) \quad (7)$$

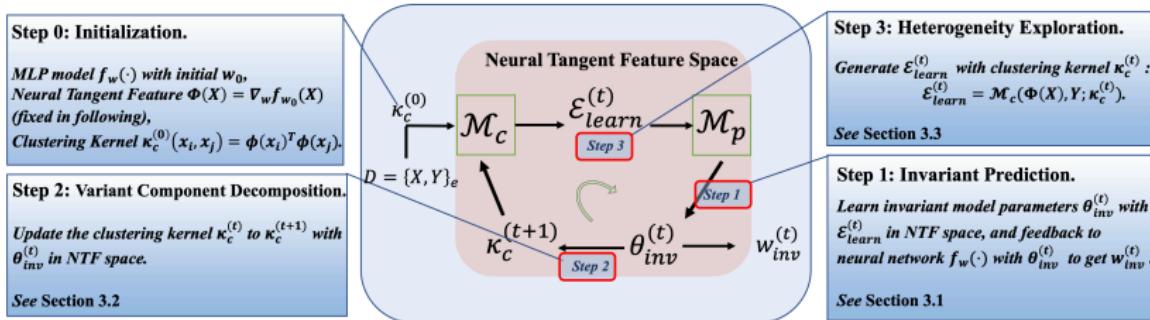
- Restrict the gradient across environments to be the same.
- Only use invariant features.
- For feature selection, we adopt the continuous feature selection method that allows for continuous optimization of M :

$$\mathcal{L}^e(\theta, \mu) = \mathbb{E}_{P^e} \mathbb{E}_M [\ell(M \odot X^e, Y^e; \theta) + \alpha \|M\|_0] \quad (8)$$

- $\|M\|_0$ controls the number of selected features.
- Conduct continuous optimization as⁹.

⁸Koyama, M., & Yamaguchi, S. (2021). When is invariance useful in an Out-of-Distribution Generalization problem?

⁹Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. Feature selection using stochastic gates, in ICML2020

Empirical Algorithm 2: Kernelized Heterogeneous Risk Minimization(KerHRM¹⁰)

• Step 0:

$$f_w(X) \approx f_{w_0}(X) + \nabla_w f_{w_0}(X)^T (w - w_0) \quad (9)$$

$$= f_{w_0}(X) + \Phi(X)^T (w - w_0) \quad (10)$$

$$\approx f_{w_0}(X) + USV^T (w - w_0) \quad (11)$$

$$= f_{w_0}(X) + \Psi(X) (\nabla_w f_{w_0}(X)(w - w_0)) = f_{w_0}(X) + \Psi(X)\theta \quad (12)$$

where $\Psi(X) \in \mathbb{R}^k$ is called the reduced Neural Tangent Features (Reduced NTFs), which convert the complicated data, non-linear setting into raw feature data, linear setting.

¹⁰ Jiashuo Liu, Zheyuan Hu, Peng Cui et al. Kernelized Heterogeneous Risk Minimization. In NeurIPS 2021.

Algorithms

- Step 1: \mathcal{M}_p Invariant Learning with Reduced NTFs $\Psi(X)$ ¹¹:

$$\theta_{inv} = \arg \min_{\theta} \sum_{e \in \mathcal{E}_{learn}} \mathcal{L}^e(\theta; \Psi, Y) + \alpha \text{Var}_{\mathcal{E}_{learn}}(\nabla_{\theta} \mathcal{L}^e) \quad (13)$$

The obtained θ_{inv} captures the invariant component in data, which can be used to wipe out the invariant part inside data.

- Step 2: Variant Component Decomposition with θ_{inv} .

- The initial similarity of two data points x_i and x_j :

$$\kappa_c^{(0)}(x_i, x_j) = \phi(x_i)^T \phi(x_j) = \langle U_i S, U_j S \rangle \quad (14)$$

- Wipe out the invariant component with θ_{inv} :

$$\Psi_V^{(t+1)}(x_i) \leftarrow U_i S - \left\langle U_i S, \theta_{inv}^{(t)} \right\rangle \theta_{inv}^{(t)} / \|\theta_{inv}^{(t)}\|^2 \quad (15)$$

- Obtain a new kernel for clustering:

$$\kappa_c^{(t+1)}(x_i, x_j) = \Psi_V^{(t+1)}(x_i)^T \Psi_V^{(t+1)}(x_j) \quad (16)$$

¹¹Here we adopt the regularizer proposed in 'Masanori Koyama, Shoichiro Yamaguchi. When is invariance useful in an Out-of-Distribution Generalization problem ?'

Algorithms

- Step 3: \mathcal{M}_c Heterogeneity Exploration with κ_c

- Capture the different relationship between Ψ_V^* and Y .
- Use $P(Y|\Psi_V)$ as the cluster centre: assume the j -th cluster centre $P_{\Theta_j}(Y|\Psi_V(X))$ to be a Gaussian around $f(\Theta_j; \Psi_V(X))$ as:

$$h_j(\Psi_V(X), Y) = P_{\Theta_j}(Y|\Psi_V(X)) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(Y - f(\Theta_j; \Psi_V(X)))^2 / 2\sigma^2) \quad (17)$$

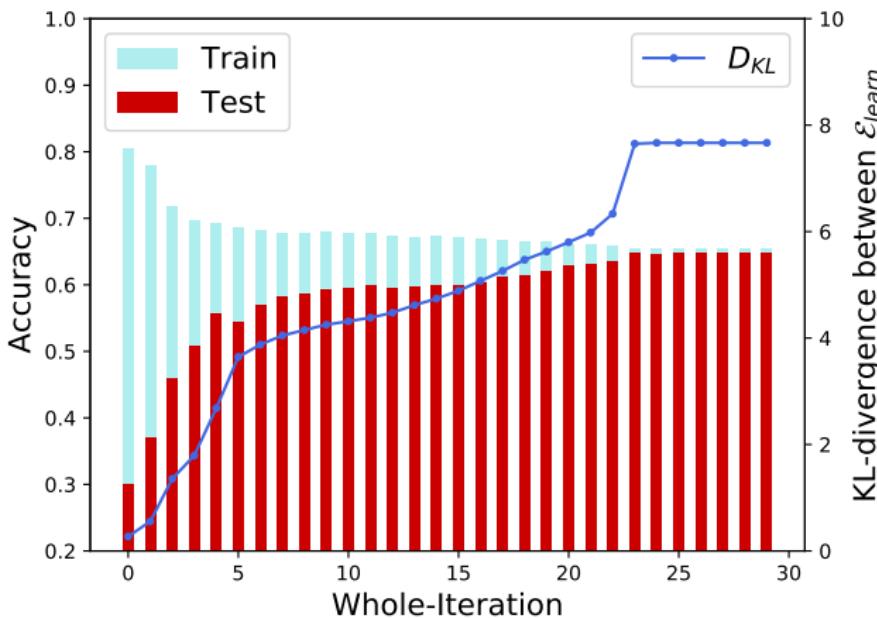
- Propose on convex clustering algorithm, which finds a mixture distribution in distribution set \mathcal{Q} defined as:

$$\mathcal{Q} = \{Q : Q = \sum_{k \in [K]} q_j h_j\} \quad (18)$$

and gives the objective function:

$$\min_{Q \in \mathcal{Q}} D_{KL}(\hat{P}_N \| Q) \Leftrightarrow \min_{\Theta, q} \left\{ \mathcal{L}_c = -\frac{1}{N} \sum_{i \in [N]} \log \left[\sum_{j \in [K]} q_j h_j(\psi_V(x_i), y_i) \right] \right\} \quad (19)$$

Surprising Results



- D_{KL} denotes $KL(P_1(Y|C) \| P_2(Y|C))$

- ① Causality & Invariance & Heterogeneity
- ② Invariant Learning Problem under Latent Heterogeneity
- ③ Distributional Invariance Property

α_0 -Distributional Invariance Property¹²

Theoretical drawbacks of the invariant learning under latent heterogeneity:

- **Strict invariance should be relaxed.**
 - The environment learning/splitting process is likely to **violate** the underlying strict invariance property.
 - If we still pursue the **strict invariance** that $Y \perp \mathcal{E}_{\text{learn}} | \Phi$, we may only obtain random noises.
- **'Invariance to what' should be characterized.**
 - The properties of $\mathcal{E}_{\text{learn}}$ are vague.
 - Cannot explain to what the learned model is invariant.

We propose the α_0 -**Distributional Invariance** to address:

- To what extent the invariance holds: we allow for some violations on the relationship $\Phi \rightarrow Y$
- To what the invariance is considered: we only consider sub-populations larger than ratio α_0

¹² Jiashuo Liu, Jiayun Wu, et al. Distributionally Invariant Learning: Rationalization and Practical Algorithms.(Working paper)

α -Distributional Invariance Property

Definition (α_0 -Distributional Invariance Property)

Given observed data distribution $P_0(X, Y)$ with latent heterogeneity, assume a lower bound $\alpha_0 \in (0, \frac{1}{2})$ on the sub-population proportion α and consider the set of potential minority sub-populations

$$\mathcal{P}_{\alpha_0}(P_0) = \{Q : P_0 = \alpha Q + (1 - \alpha)Q_0, \text{ for } \alpha \in [\alpha_0, 1] \text{ and distribution } Q_0 \ll^{13} P_0\} \quad (20)$$

Then a representation Φ is α_0 -distributionally invariant if

$$\underbrace{\sup_{Q \in \mathcal{P}_{\alpha_0}(P_0(X, Y))} \rho(Q(Y|\Phi), P_0(Y|\Phi))}_{\text{to what it is invariant}} \leq \underbrace{\delta}_{\text{to what extent}} \quad \text{with some } \delta > 0 \quad (21)$$

where $\rho(\cdot, \cdot)$ is some distance metric between two distributions (e.g., MMD distance, KL divergence). For simplicity, for representation Φ that is α_0 -distributionally invariant, we denote it as $Y \perp^{\delta} \mathcal{E}_{\alpha_0}(P_0)|\Phi$, where $\mathcal{E}_{\alpha_0}(P_0)$ denotes the random variable on indices of distributions in $\mathcal{P}_{\alpha_0}(P_0)$.

- Based on this, we propose the Distributionally Invariant Learning (DIL) framework¹⁴.
- We could derive the generalization error bound for our method.

¹³ $Q_0 \ll P_0$ means the support of Q_0 is no larger than P_0

¹⁴ Jiashuo Liu, Jiayun Wu, et al. Distributionally Invariant Learning: Rationalization and Practical Algorithms.
<https://arxiv.org/abs/2206.02990>

Relationship with Traditional Invariance

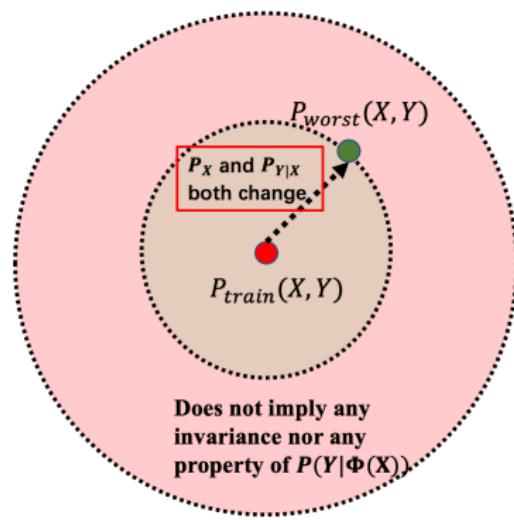
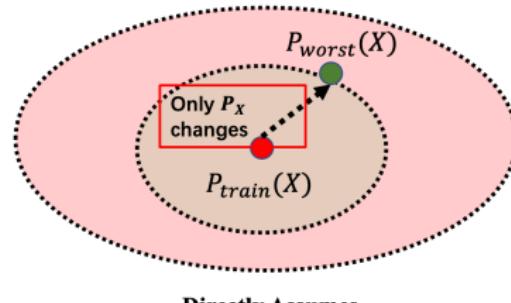
- With Invariance Property in IRM:

- set $\delta = 0$
- set $\mathcal{P}_{\alpha_0} = \text{supp}(\mathcal{E})$
- set $\rho(P, Q) = \mathbb{E}[P] - \mathbb{E}[Q]$

- With Invariance Property in MIP:

- set $\delta = 0$
- set $\mathcal{P}_{\alpha_0} = \text{supp}(\mathcal{E})$
- set $\rho(P, Q) = \text{MMD}(P, Q)$ or $\rho(P, Q) = \text{KL}(P\|Q)$

Relationship with DRO

**Joint DRO****Marginal DRO**

Relationship with DRO

