Data Heterogeneity & Invariance in Out-of-Distribution Generalization

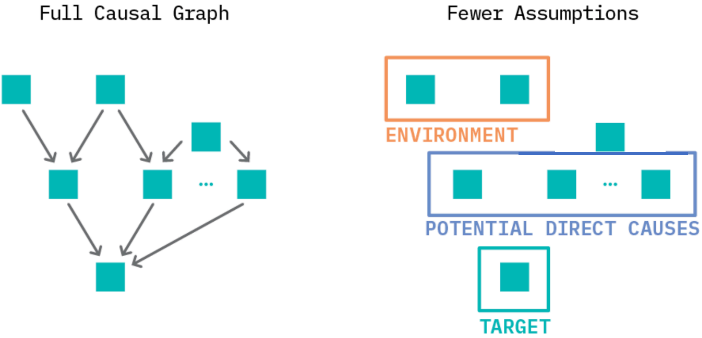
Jiashuo Liu

Tsinghua University

2023.02.18

1. Causality & Invariance
2. Invariance & Heterogeneity
3. Invariant Learning Problem under Latent Heterogeneity
4. Distributional Stability

Causality & Invariance



[1]

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

Invariance Property

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

Assumption (Invariance Assumption[2])

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

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

- *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 Assumption [3])

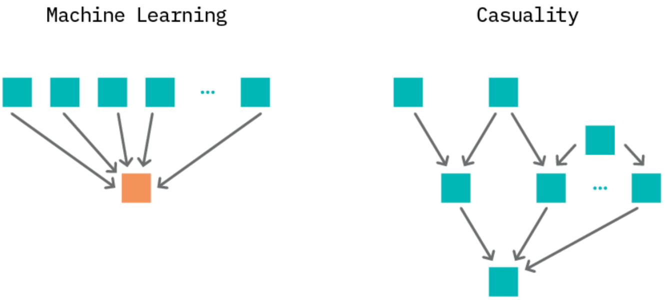
There exists random variable $\Phi(X)$ such that for all $e_1, e_2 \in \mathrm{supp}(\mathcal{E})$, we have

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

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

[3] Arjovsky, 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



For prediction problem, the invariance property is enough (we only care about $\mathrm{Pa}(Y)$). $\Rightarrow$ 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 optimality[4].

[4] Koyama, 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:

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

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

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

Invariant Risk Minimization [5]

- Idea: learn an invariant predictor Φ with invariant $P(Y|\Phi, e)$ for $e \in \text{supp}(\mathcal{E})$

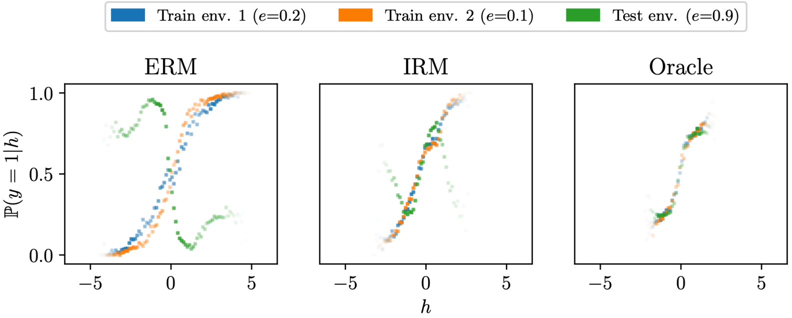
$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H} \\ w:\mathcal{H}\to\mathcal{Y}}} \quad \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(w\circ\Phi)$$
$$\text{subject to} \quad w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w}\circ\Phi), \text{ for all } e\in\mathcal{E}_{\text{tr}}.$$

- Approximation:

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(\Phi) + \lambda \cdot \|\nabla_{w|w=1.0}\, R^e(w\cdot\Phi)\|^2, \qquad \text{(IRMv1)}$$

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

Invariant Risk Minimization [6]



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

Following Works

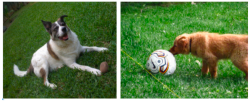
- S. Chang, Y. Zhang, M. Yu, and T. S. Jaakkola, Invariant rationalization, in ICML, ser. PMLR, vol. 119. PMLR, 2020, pp. 14481458.
- M. Koyama and S. Yamaguchi, Out-of-distribution generalization with maximal invariant predictor, CoRR, vol. abs/2008.01883, 2020. [Online]. Available: https://arxiv.org/abs/2008.01883
- K. Ahuja, K. Shanmugam, K. Varshney, and A. Dhurandhar, Invariant risk minimization games, in ICML. PMLR, 2020, pp. 145155.
- D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, Out-of-distribution generalization via risk extrapolation (rex), in ICML. PMLR, 2021, pp. 58155826.
- D. Mahajan, S. Tople, and A. Sharma, Domain generalization using causal matching, in ICML, ser. PMLR, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 1824 Jul 2021, pp. 73137324.
- E. Rosenfeld, P. Ravikumar, and A. Risteski, The risks of invariant risk minimization, arXiv preprint arXiv:2010.05761, 2020.
- P. Kamath, A. Tangella, D. Sutherland, and N. Srebro, Does invariant risk minimization capture invariance? in AISTATS. PMLR, 2021, pp. 40694077.
- K. Ahuja, J. Wang, A. Dhurandhar, K. Shanmugam, and K. R. Varshney, Empirical or invariant risk minimization? a sample complexity perspective, arXiv preprint arXiv:2010.16412, 2020.

1. Causality & Invariance
2. Invariance & Heterogeneity
3. Invariant Learning Problem under Latent Heterogeneity
4. Distributional Stability

Invariance Property

Invariance to What? $\Rightarrow$ Some limitations in practice.

Env1



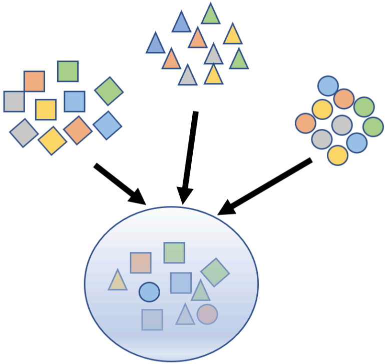
Env2

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.



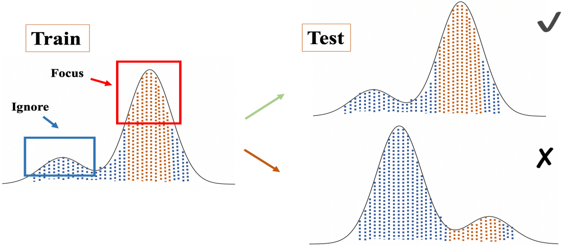
Limitation 2: Quality of environments

- 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.) [7]

[7] Causality for Machine Learning. Cloudera's Fast Forward Labs

Measure the Predictive Heterogeneity[8]

- Idea: measure the heterogeneity inside data via information gain

Definition 3 (Interaction Heterogeneity). *Let X, Y be random variables taking values in $\mathcal{X} \times \mathcal{Y}$. Denote the set of random categorical variables as $\mathcal{C}$, and take its subset $\mathscr{E} \subseteq \mathcal{C}$. Then $\mathscr{E}$ is an environment set iff there exists $\mathcal{E} \in \mathscr{E}$ such that $X, Y \perp\!\!\!\perp \mathcal{E}$. $\mathcal{E} \in \mathscr{E}$ is called an environment variable. The interaction heterogeneity between X and Y w.r.t. the environment set $\mathscr{E}$ is defined as:*

$$\mathcal{H}^{\mathscr{E}}(X,Y) = \sup_{\mathcal{E}\in\mathscr{E}} \mathbb{I}(Y;X|\mathcal{E}) - \mathbb{I}(Y;X). \tag{6}$$

Definition 4 (Conditional Predictive $\mathcal{V}$-information). *Let X, Y be two random variables taking values in $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{E}$ be an environment variable. The conditional predictive $\mathcal{V}$-information is defined as:*

$$\mathbb{I}_{\mathcal{V}}(X \to Y|\mathcal{E}) = H_{\mathcal{V}}(Y|\emptyset,\mathcal{E}) - H_{\mathcal{V}}(Y|X,\mathcal{E}), \tag{7}$$

where $H_{\mathcal{V}}(Y|\emptyset,\mathcal{E})$ and $H_{\mathcal{V}}(Y|X,\mathcal{E})$ are defined as:

$$H_{\mathcal{V}}(Y|X,\mathcal{E}) = \mathbb{E}_{e\sim\mathcal{E}}\left[\inf_{f\in\mathcal{V}} \mathbb{E}_{x,y\sim X,Y|\mathcal{E}=e}[-\log f[x](y)]\right]. \tag{8}$$

$$H_{\mathcal{V}}(Y|\emptyset,\mathcal{E}) = \mathbb{E}_{e\sim\mathcal{E}}\left[\inf_{f\in\mathcal{V}} \mathbb{E}_{y\sim Y|\mathcal{E}=e}[-\log f[\emptyset](y)]\right]. \tag{9}$$

[8] Measure the Predictive Heterogeneity. Jiashuo Liu *et al.* ICLR 2023.

Measure the Predictive Heterogeneity[9]

Definition 4 (Conditional Predictive $\mathcal{V}$-information). *Let X, Y be two random variables taking values in $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{E}$ be an environment variable. The conditional predictive $\mathcal{V}$-information is defined as:*

$$\mathbb{I}_{\mathcal{V}}(X \to Y|\mathcal{E}) = H_{\mathcal{V}}(Y|\emptyset, \mathcal{E}) - H_{\mathcal{V}}(Y|X, \mathcal{E}), \tag{7}$$

where $H_{\mathcal{V}}(Y|\emptyset, \mathcal{E})$ and $H_{\mathcal{V}}(Y|X, \mathcal{E})$ are defined as:

$$H_{\mathcal{V}}(Y|X, \mathcal{E}) = \mathbb{E}_{e\sim\mathcal{E}}\left[\inf_{f\in\mathcal{V}} \mathbb{E}_{x,y\sim X,Y|\mathcal{E}=e}[-\log f[x](y)]\right]. \tag{8}$$

$$H_{\mathcal{V}}(Y|\emptyset, \mathcal{E}) = \mathbb{E}_{e\sim\mathcal{E}}\left[\inf_{f\in\mathcal{V}} \mathbb{E}_{y\sim Y|\mathcal{E}=e}[-\log f[\emptyset](y)]\right]. \tag{9}$$

Definition 5 (Predictive Heterogeneity). *Let X, Y be random variables taking values in $\mathcal{X} \times \mathcal{Y}$ and $\mathscr{E}$ be an environment set. The predictive heterogeneity for the prediction $X \to Y$ with respect to $\mathscr{E}$ is defined as:*

$$\mathcal{H}_{\mathcal{V}}^{\mathscr{E}}(X \to Y) = \sup_{\mathcal{E}\in\mathscr{E}} \mathbb{I}_{\mathcal{V}}(X \to Y|\mathcal{E}) - \mathbb{I}_{\mathcal{V}}(X \to Y), \tag{10}$$

where $\mathbb{I}_{\mathcal{V}}(X \to Y)$ is the predictive $\mathcal{V}$-information following from Definition 2.

[9] Measure the Predictive Heterogeneity. Jiashuo Liu *et al.* ICLR 2023.

Measure the Predictive Heterogeneity[10]

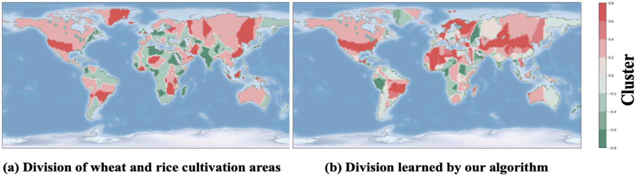
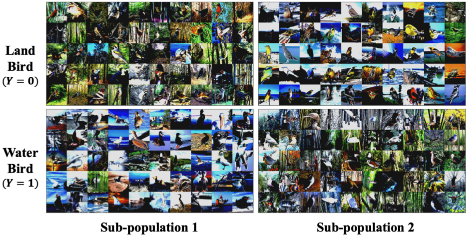
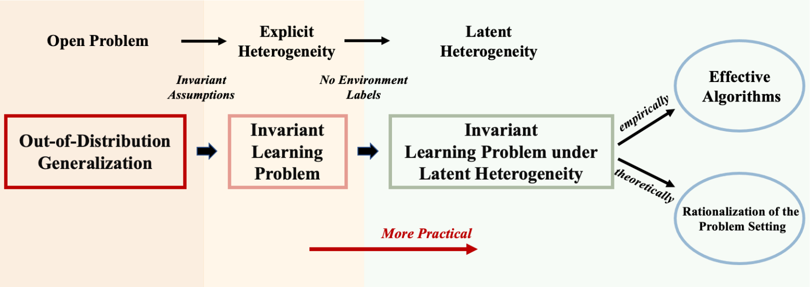


Figure 1: Results on the crop yield data. We color each region according to its main crop type, and the shade represents the proportion of the main crop type after smoothing via k-means ($k = 3$).



[10] Measure the Predictive Heterogeneity. Jiashuo Liu *et al.* ICLR 2023.

An Overview


Open Problem
Explicit Heterogeneity
Latent Heterogeneity
Invariant Assumptions
No Environment Labels
Out-of-Distribution Generalization
Invariant Learning Problem
Invariant Learning Problem under Latent Heterogeneity
empirically
theoretically
Effective Algorithms
Rationalization of the Problem Setting
More Practical

Invariant Learning Problem under Latent Heterogeneity

Assumption (Heterogeneity Assumption)

For random variable pair (X, Φ^*) and Φ^* satisfying the Invariance Assumption, using functional representation lemma[11], there exists random variable Ψ^* such that $X = X(\Phi^*, \Psi^*)$, then we assume $P^e(Y|\Psi^*)$ can arbitrary change across environments $e \in \mathrm{supp}(\mathcal{E})$.

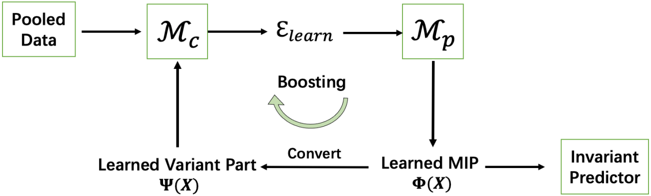
Problem (Invariant Learning Problem under Latent Heterogeneity)

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

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

Empirical Algorithm 1: Heterogeneous Risk Minimization[12]

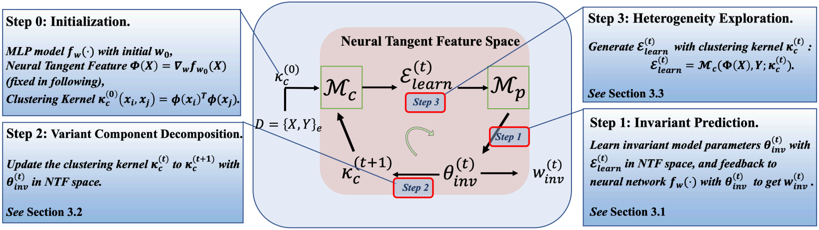
- This work temporarily focuses on a simple but general setting, where $X = [\Phi^*, \Psi^*]^T$ at the raw feature level.
- The HRM framework contains two modules, named **Heterogeneity Identification** module $\mathcal{M}_c$ and **Invariant Prediction** module $\mathcal{M}_p$.



- The two modules can **mutually promote** each other, meaning that the invariant prediction and the quality of $\mathcal{E}_{learn}$ 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 $\Psi(X) = (1 - M) \odot X$.

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

Empirical Algorithm 2: Kernelized Heterogeneous Risk Minimization(KerHRM[13])



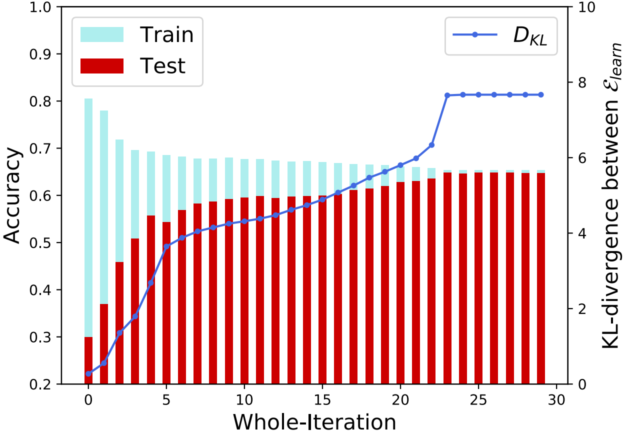
- Step 0:

$$
\begin{aligned}
f_w(X) &\approx f_{w_0}(X) + \nabla_w f_{w_0}(X)^T(w - w_0) && (5)\\
&= f_{w_0}(X) + \Phi(X)^T(w - w_0) && (6)\\
&\approx f_{w_0}(X) + USV^T(w - w_0) && (7)\\
&= f_{w_0}(X) + \Psi(X)\left(V^T(w - w_0)\right) = f_{w_0}(X) + \Psi(X)\theta && (8)
\end{aligned}
$$

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.

[13] Jiashuo Liu, Zheyuan Hu, Peng Cui *et al.* Kernelized Heterogeneous Risk Minimization. *In NeurIPS 2021*.

Surprising Results



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

Measure the Stability

- Measure via Directional Worst-Case[14]
  - Sign Stability:

$$\begin{aligned} s = \exp(-\inf_{Q} D_{KL}(Q \| P_{tr})) \\ \text{s.t.} \quad \text{sign}(\theta(Q)) \neq \text{sign}(\theta(P_{tr})) \end{aligned} \tag{9}$$

  - Beyond Omni-Directional:

$$\begin{aligned} s = \exp(-\inf_{Q:Q(\cdot|E)=P_{tr}(\cdot|E)} D_{KL}(Q \| P_{tr})) \\ \text{s.t.} \quad \text{sign}(\theta(Q)) \neq \text{sign}(\theta(P_{tr})) \end{aligned} \tag{10}$$

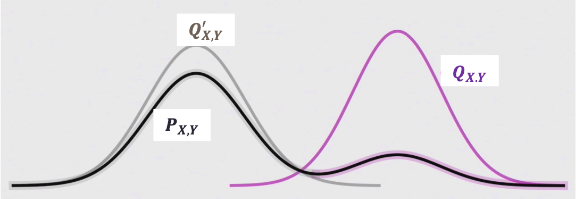
    which only considers the shifts on $Q(E)$.
- Measure via Prediction Risk[15]

$$I_r(P) := \inf_{Q} \left\{ D_{KL}(Q \| P_{tr}) : \mathbb{E}_Q[R] \geq r \right\} \tag{11}$$

[14] Distributionally robust and generalizable inference. Dominik Rothenhäusler, Peter Bühlmann.
[15] Namkoong *et al.* Minimax Optimal Estimation of Stability Under Distribution Shift.

Sub-population

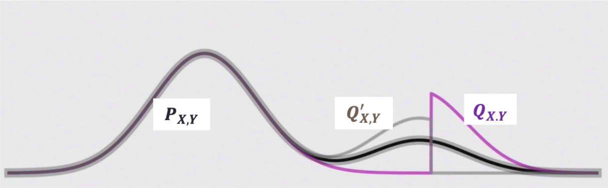
$Q_{X,Y}$ is a **subpopulation** $\Longleftrightarrow$ $\exists$proportion $\alpha \in (0,1]$, prob. $Q'_{X,Y}$, s.t. $P_{X,Y}(\cdot) = \alpha Q_{X,Y} + (1-\alpha)Q'_{X,Y}$



Figures from Namkoong's talk : https://drive.google.com/file/d/1ApBFWEkzOP39gIVMDBnXbdXXlARWXqDX/view

Sub-population

$\boldsymbol{Q_{X,Y}}$ is a **subpopulation** $\Longleftrightarrow$ $\exists$proportion $\alpha \in (0,1]$, prob. $Q'_{X,Y}$, s.t. $P_{X,Y}(\cdot) = \alpha \boldsymbol{Q_{X,Y}} + (1-\alpha) Q'_{X,Y}$



Figures from Namkoong's talk : https://drive.google.com/file/d/1ApBFWEkzOP39gIVMDBnXbdXXlARWXqDX/view

Distributional Stability

Automatically find **worst-subpopulations** and measure the **discrepancy** on the distribution of $\boldsymbol{Y|X}$

$Q_{X,Y} \succcurlyeq \alpha_0 \iff$ subpopulation with proportion larger than $\alpha_0 \in (0, ½)$

α_0-Distributional Stability, DS_{α_0}

$$DS_{\alpha_0}(X \to Y; P_{tr}) := \sup_{Q_{X,Y} \succcurlyeq \alpha_0} \rho\big(Q(Y|X), P_{tr}(Y|X)\big)$$

where $\rho(\cdot,\cdot)$ is distribution distance metric.

Relationship with Strict Invariance

α_0-Distributional Stability, DS_{α_0}

$$DS_{\alpha_0}(X \to Y; P_{tr}) := \sup_{Q_{X,Y} \succcurlyeq \alpha_0} \rho\big(Q(Y|X), P_{tr}(Y|X)\big)$$

where $\rho(\cdot,\cdot)$ is distribution distance metric.

$\rho\big(Q(Y|X), P(Y|X)\big)$
$= \mathbb{E}[|\mathbb{E}_Q[Y|X] - \mathbb{E}_P[Y|X]|^2]$

$\downarrow$

$\rho(\cdot,\cdot) = D_{KL}(Q(Y|X)||P(Y|X))$ or
$\rho(\cdot,\cdot) = MMD\ (Q(Y|X), P(Y|X))$

Strict Invariance

***form 1*: for any $e_i, e_j \in$ supp$(\mathcal{E})$,**

$$\mathbb{E}[Y|X, e_i] = \mathbb{E}[Y|X, e_j]$$

***form 2*: for any $e_i, e_j \in$ supp$(\mathcal{E})$,**

$$P(Y|X, e_i) = P(Y|X, e_j)$$

OOD Generalization Regret Bounds

Assume the problem is learnable *w.r.t.* an expansion function $s(\cdot)$, and choose $\rho(\cdot,\cdot)$ as KL-divergence. Then for $\Phi \in \Upsilon$, we have:

$$\mathbb{E}_{P_{test}}\boxed{\left[\left|\mathbb{E}_{P_{test}}[\ell(f(\Phi))|\Phi] - \mathbb{E}_{P_{train}}[\ell(f(\Phi))|\Phi]\right|\right]} \le \mathcal{O}\left(\sqrt{s\left(DS_{\alpha_0}(\Phi \to Y; P_{train})\right)}\right)$$

regret on the testing distribution

distributional stability

Contact

Jiashuo Liu

- (+86) 13015155336
- @liujiashuo77
- liujiashuo77@gmail.com
- ljsthu.github.io
- https://github.com/LJSthu