Heterogeneous Risk Minimization

International Conference on Machine Learning 2021

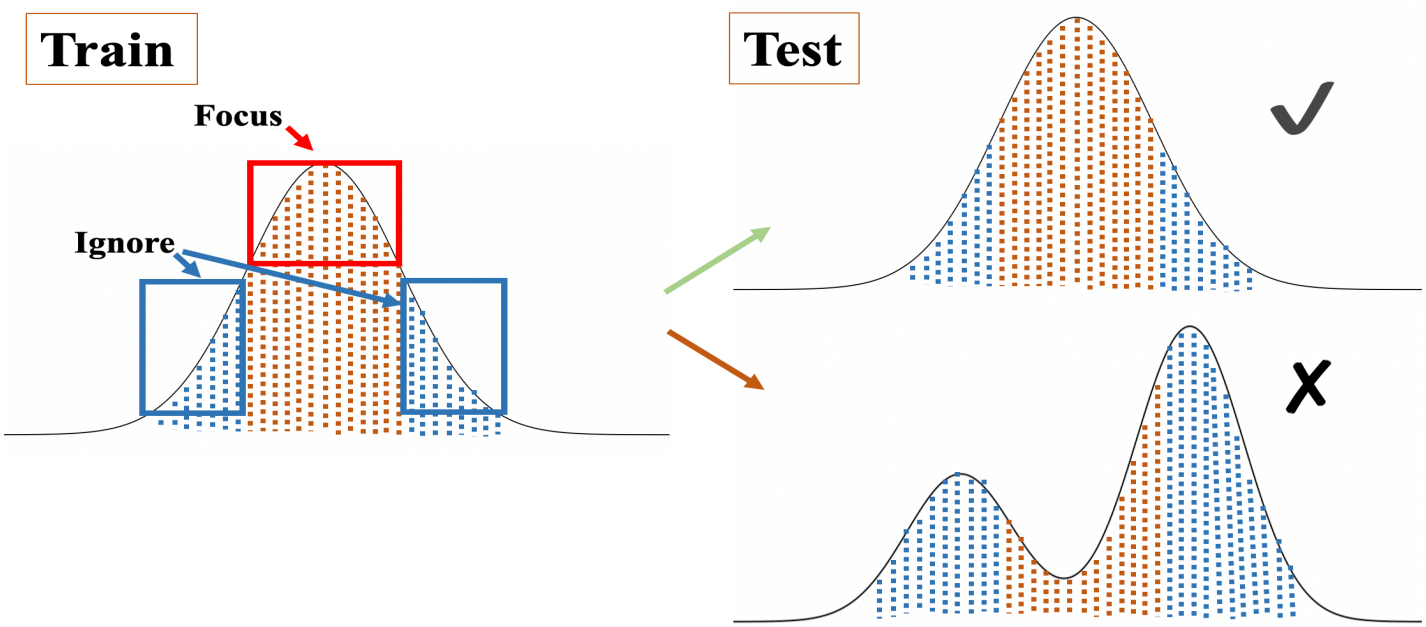
Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, Zheyan Shen

Department of Computer Science and Technology, Tsinghua University

Empirical Risk Minimization(ERM)

$$\theta_{ERM} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell(\theta; X_i, Y_i) \tag{1}$$

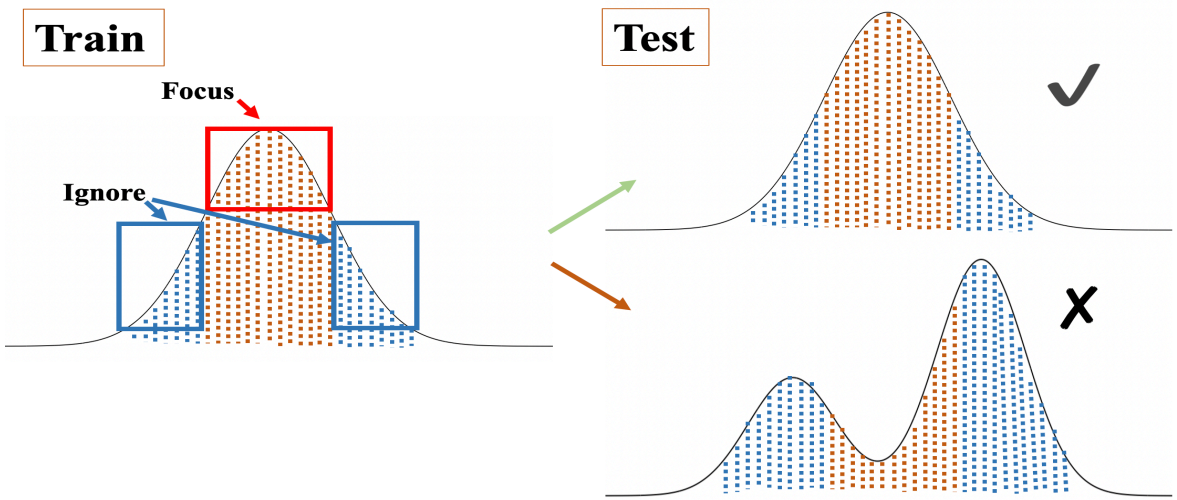
- Optimize the **average error** oof data points.
- Focus on the **major group** of data.
- Ignore the **minor group** of data $\rightarrow$ Break down under distributional shifts



Latent Heterogeneity in Data

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 Heterogeneous Risk Minimization.

Out-of-Distribution Generalization Problem(OOD 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:

$$\theta_{OOD} = \arg\min_{\theta} \max_{e \in \mathrm{supp}(\mathcal{E})} \mathcal{L}^e(\theta; X, Y) \tag{2}$$

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:

$$\mathcal{L}^e(\theta; X, Y) = \mathbb{E}_{X,Y \sim P^e}[\ell(\theta; X, Y)] \tag{3}$$

- OOD problem hopes to optimize the worst-case risk of all possible environments or distributions in $\mathrm{supp}(\mathcal{E})$

Invariance Assumption and MIP

Assumption (Invariance Assumption)

There exists random variable $\Phi^*(X)$ such that the following properties hold:

1. Invariance property*:* *for all $e_1, e_2 \in \mathrm{supp}(\mathcal{E})$, we have*

$$P^{e_1}(Y|\Phi^*(X)) = P^{e_2}(Y|\Phi^*(X)) \tag{4}$$

2. Sufficiency property*:* $Y = f(\Phi^*) + \epsilon,\ \epsilon \perp X$.

To obtain the invariant predictor $\Phi^*(X)$, one can seeks for the **Maximal Invariant Predictor**[1][2], which is defined as follows:

Definition (Invariance Set & Maximal Invariant Predictor)

The invariance set $\mathcal{I}$ with respect to $\mathcal{E}$ is defined as:

$$\mathcal{I}_{\mathcal{E}} = \{\Phi(X) : Y \perp \mathcal{E}|\Phi(X)\} = \{\Phi(X) : H[Y|\Phi(X)] = H[Y|\Phi(X), \mathcal{E}]\} \tag{5}$$

where $H[\cdot]$ is the Shannon entropy of a random variable. The corresponding maximal invariant predictor (MIP) of $\mathcal{I}_{\mathcal{E}}$ is defined as:

$$S = \arg\max_{\Phi \in \mathcal{I}_{\mathcal{E}}} I(Y; \Phi) \tag{6}$$

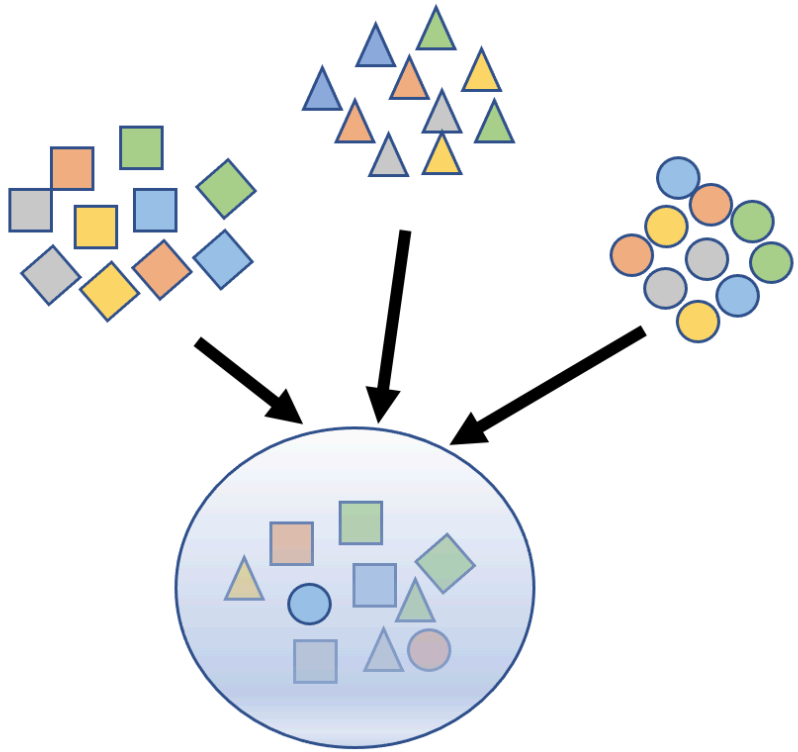
where $I(\cdot;\cdot)$ measures Shannon mutual information between two random variables.

[1] Chang, S., Zhang, Y. et al. (2020, November). Invariant rationalization.

[2] Koyama, M., & Yamaguchi, S. When is invariance useful in an Out-of-Distribution Generalization problem ?

No Training Environments

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.



Quality of Training Environments

- The flow of Invariant Learning methods:

  Given $\mathcal{E}_{tr}$ $\rightarrow$ Find MIP Φ^*_{tr} of $\mathcal{I}_{\mathcal{E}_{tr}}$ $\rightarrow$ Predict using Φ^*_{tr} $\rightarrow$ OOD "Optimal?"

- Recall the definition of MIP:

$$\arg\max_{\Phi \in \mathcal{I}_{\mathcal{E}}} I(Y; \Phi) \tag{7}$$

1. MIP relies on the invariance set $\mathcal{I}_{\mathcal{E}}$
2. Invariance set $\mathcal{I}_{\mathcal{E}}$ relies on the given environments $\mathcal{E}$.

- What happens when $\mathcal{E}$ is replaced by $\mathcal{E}_{tr}$?
  1. $\mathrm{supp}(\mathcal{E}_{tr}) \subset \mathrm{supp}(\mathcal{E})$
  2. $\mathcal{I}_{\mathcal{E}} \subset \mathcal{I}_{\mathcal{E}_{tr}}$
  3. Φ^*_{tr} NOT INVARIANT.

Remark: We need training environments where $\mathcal{I}_{\mathcal{E}_{tr}} \rightarrow \mathcal{I}_{\mathcal{E}}$

HRM Problem

Assumption (Heterogeneity Assumption)

For random variable pair (X, Φ^*) and Φ^* satisfying the Invariance Assumption, using functional representation lemma[3], 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 (Heterogeneous Risk Minimization Problem)

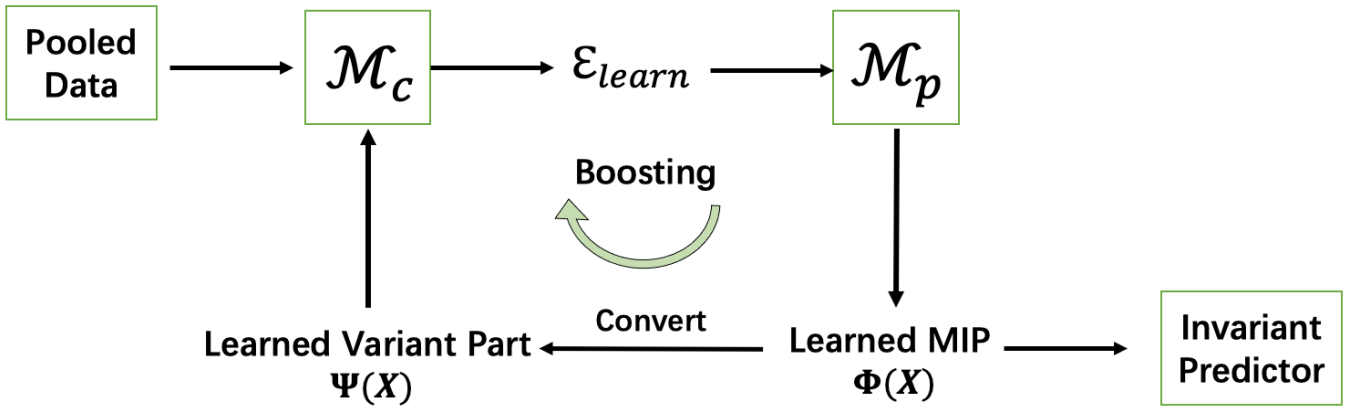
Given heterogeneous dataset $D = \{D^e\}_{e \in \mathrm{supp}(\mathcal{E}_{latent})}$ without environment labels, the task is to generate environments $\mathcal{E}_{learn}$ with minimal $|\mathcal{I}_{\mathcal{E}_{learn}}|$ and learn invariant model under learned $\mathcal{E}_{learn}$ with good OOD performance.

• This work temporarily focuses on a simple but general setting, where $X = [\Phi^*, \Psi^*]^T$ in raw feature level and Φ^*, Ψ^* satisfy the Invariance Assumption.

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

The Whole Algorithm

Our HRM 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 $\Psi(X) = (1 - M) \odot X$.

Baselines & Evaluation Criterions

Baselines:

- Empirical Risk Minimization(ERM): $\min_\theta \mathbb{E}_{P_0}[\ell(\theta; X, Y)]$
- Distributionally Robust Optimization(DRO[1]): $\min_\theta \sup_{Q \in W(Q, P_0) \leq \rho} \mathbb{E}_Q[\ell(\theta; X, Y)]$
- Environment Inference for Invariant Learning(EIIL[2]):

$$\min_{\Phi} \max_{u} \sum_{e \in \mathcal{E}} \frac{1}{N_e} \sum_i u_i(e) \ell(w \odot \Phi(x_i), y_i) + \sum_{e \in \mathcal{E}} \lambda \| \nabla_{w|w=1.0} \frac{1}{N_e} \sum_i u_i(e) \ell(w \odot \Phi(x_i), y_i) \|_2 \tag{8}$$

- Invariant Risk Minimization(IRM[3]) with environment $\mathcal{E}_{tr}$ labels:

$$\min_{\Phi} \sum_{e \in \mathcal{E}_{tr}} \mathcal{L}^e + \lambda \| \nabla_{w|w=1.0} \mathcal{L}^e(w \odot \Phi) \|^2 \tag{9}$$

Evaluation Criterion:

- Mean_Error: $\text{Mean\_Error} = \frac{1}{|\mathcal{E}_{test}|} \sum_{e \in \mathcal{E}_{test}} \mathcal{L}^e$
- Std_Error: $\text{Std\_Error} = \sqrt{\frac{1}{|\mathcal{E}_{test}|-1} \sum_{e \in \mathcal{E}_{test}} (\mathcal{L}^e - \text{Mean\_Error})^2}$
- Max_Error: $\text{Max\_Error} = \max_{e \in \mathcal{E}_{test}} \mathcal{L}^e$

Selection Bias

• Setting: $X = [\Phi^*, \Psi^*]^T \in \mathbb{R}^d$ and $Y = f(\Phi^*) + \epsilon$ and that $P(Y|\Phi^*)$ remains invariant across environments while $P(Y|\Psi^*)$ changes arbitrarily. We select data points according to a certain variable set $V_b \subset \Psi^*$:

$$\hat{P}(x) = \prod_{v_i \in V_b} |r|^{-5*|f(\phi^*) - sign(r)*v_i|} \tag{10}$$

where $|r| > 1$, $V_b \in \mathbb{R}^{n_b}$ and $\hat{P}(x)$ denotes the probability of point x to be selected.

- Training: $sum = 2000$ data points, where $\kappa = 95\%$ points from environment e_1 with a predefined r and $1 - \kappa = 5\%$ points from e_2 with $r = -1.1$.
- Testing: 10 environments with $r \in [-3, -2.7, -2.3, \ldots, 2.3, 2.7, 3.0]$.

Some demonstrations:

- $|r|$ eventually controls the strengths of the spurious correlation between V_b and Y, the larger $|r|$, the more biased the data are.
- $\mathrm{sign}(r)$ controls the direction of the spurious correlation between V_b and Y.

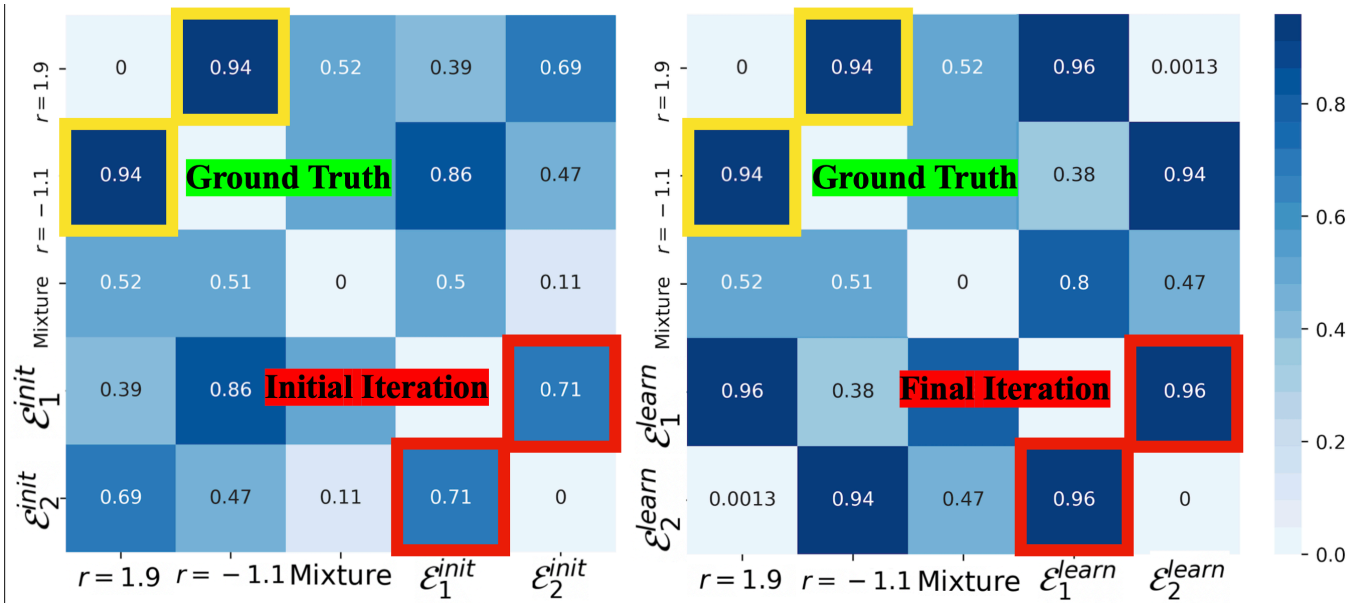
Selection Bias Results

表 1: Results in selection bias simulation experiments of different methods with varying selection bias r, and dimensions n_b and d of training data, and each result is averaged over ten times runs.

Scenario 1: varying selection bias rate r $(d = 10, n_b = 1)$									
r	$r = 1.5$			$r = 1.9$			$r = 2.3$		
Methods	Mean_Error	Std_Error	Max_Error	Mean_Error	Std_Error	Max_Error	Mean_Error	Std_Error	Max_Error
ERM	0.476	0.064	0.524	0.510	0.108	0.608	0.532	0.139	0.690
DRO	0.467	0.046	0.516	0.512	0.111	0.625	0.535	0.143	0.746
EIIL	0.477	0.057	0.543	0.507	0.102	0.613	0.540	0.139	0.683
IRM(with $\mathcal{E}_{tr}$ label)	0.460	0.014	0.475	0.456	0.015	0.472	0.461	0.015	0.475
HRM^s	0.465	0.045	0.511	0.488	0.078	0.577	0.506	0.096	0.596
HRM	**0.447**	**0.011**	**0.462**	**0.449**	**0.010**	**0.465**	**0.447**	**0.011**	**0.463**
Scenario 2: varying dimension d $(r = 1.9, n_b = 0.1d)$									
d	$d = 10$			$d = 20$			$d = 40$		
Methods	Mean_Error	Std_Error	Max_Error	Mean_Error	Std_Error	Max_Error	Mean_Error	Std_Error	Max_Error
ERM	0.510	0.108	0.608	0.533	0.141	0.733	0.528	0.175	0.719
DRO	0.512	0.111	0.625	0.564	0.186	0.746	0.555	0.196	0.758
EIIL	0.507	0.102	0.613	0.543	0.147	0.699	0.542	0.178	0.727
IRM(with $\mathcal{E}_{tr}$ label)	0.456	0.015	0.472	0.484	0.014	0.489	0.500	0.051	0.540
HRM^s	0.488	0.078	0.577	0.486	0.069	0.555	0.477	0.081	0.553
HRM	**0.449**	**0.010**	**0.465**	**0.466**	**0.011**	**0.478**	**0.465**	**0.015**	**0.482**

Selection Bias Results

We visualize the differences between environments using Task2Vec[4] as follows:



- The quality of $\mathcal{E}_{learn}$ becomes better.
- The quality of $\mathcal{E}_{learn}$ is even better than the ground truth environments.

[4] Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C. C., Soatto, S., and Perona, P. Task2vec: Task embedding for meta-learning.

Notes

Due to time limits, please refer to our paper

https://arxiv.org/pdf/2105.03818.pdf

for:

- The details of HRM framework
- The theoretical analysis of the role of environments in invariant learning
- The theoretical analysis of the mutual promotion
- More experiments, including selection bias, anti-causal effect and real data.

References

1. Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. International Conference on Learning Representations, 2018.
2. Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In ICML Workshop on Uncertainty and Robustness, 2020.
3. Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez- Paz, D. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019.