From Latent Heterogeneity to Out-of-Distribution Generalization

Heterogeneous Risk Minimization(ICML 2021)
Kernelized Heterogeneous Risk Minimization(NeurIPS 2021)

Jiashuo Liu

TrustWorthy-AI Group, CST, Tsinghua University

2021.10.18

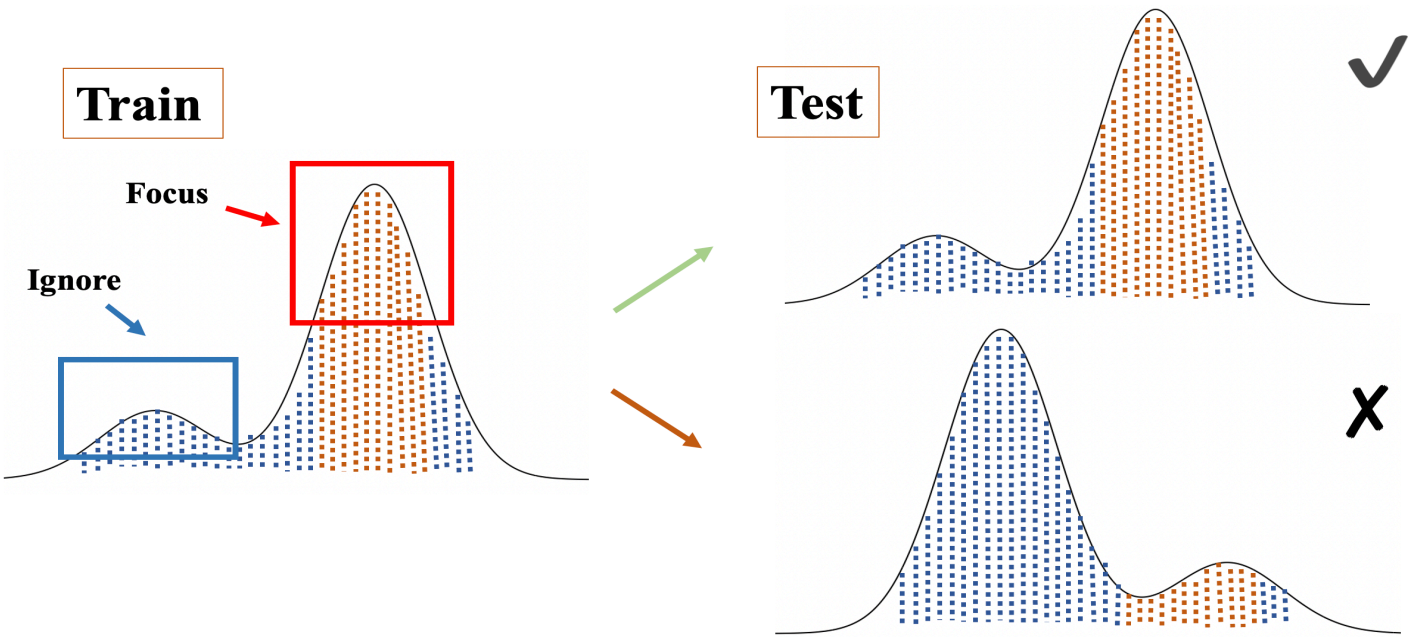
1. Background of OOD Generalization problem
2. Invariance-Based Optimization
3. Limitations
4. Heterogeneous Risk Minimization(HRM)
5. Kernelized Heterogeneous Risk Minimization(KerHRM)
6. Conclusion

1. Background of OOD Generalization problem
2. Invariance-Based Optimization
3. Limitations
4. Heterogeneous Risk Minimization(HRM)
5. Kernelized Heterogeneous Risk Minimization(KerHRM)
6. Conclusion

Empirical Risk Minimization(ERM)

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

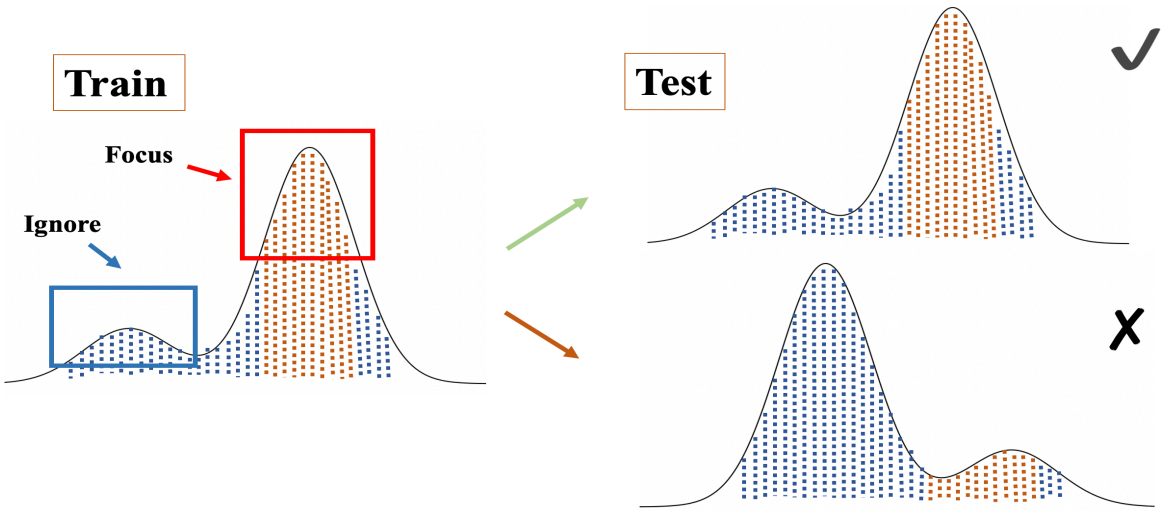
- Optimize the **average error** oof data points.
- Focus on the **major group** of data.
- Ignore the **minor group** of data $\rightarrow$ Break down under distributional shifts



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 $\mathrm{supp}(\mathcal{E})$

Related Works

$$f_\theta^* = \arg\min_{f_\theta} \mathbb{E}_{X,Y\sim P_{tr}}[\ell(f_\theta(X), Y)] \tag{4}$$

Categorize the existing methods into three parts based on their positions in the whole learning pipeline accordingly[1]:

- **Unsupervised Representation Learning**: Disentangled Representation Learning, Causal Representation Learning
- **Supervised Model Learning**: Causal Learning, Stable Learning, Domain Generalization
- **Optimization**: Distributionally Robust Optimization, Invariance-Based Optimization

[1] Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., & Cui, P. (2021). Towards out-of-distribution generalization: A survey. arXiv preprint arXiv:2108.13624.
Website for paper list: http://out-of-distribution-generalization.com

1 Background of OOD Generalization problem

2 Invariance-Based Optimization

3 Limitations

4 Heterogeneous Risk Minimization(HRM)

5 Kernelized Heterogeneous Risk Minimization(KerHRM)

6 Conclusion

Invariance Assumption

To deal with the potential distributional shifts, one common assumption made in invariant learning is the **Invariance Assumption**.

Assumption (Invariance Assumption)

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

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

$$P^{e_1}(Y|\Phi^*(X)) = P^{e_2}(Y|\Phi^*(X)) \tag{5}$$

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

Here we make some demonstrations on the Invariance Assumption:

- The first property assumes that the relationship between $\Phi^*(X)$ and Y remains invariant across environments, which is also referred to as causal relationship.
- The second property assumes that $\Phi^*(X)$ can provide all information of the target label Y.
- $\Phi^*(X)$ is referred to as **(Causally) Invariant Predictors**.

Maximal Invariant Predictor

To obtain the invariant predictor $\Phi^*(X)$, one can seeks for the **Maximal Invariant Predictor**[2][3], which is defined as follows:

Definition (Invariance Set & Maximal Invariant Predictor)

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

$$\mathcal{I}_{\mathcal{E}} = \{\Phi(X) : Y \perp \mathcal{E} | \Phi(X)\} = \{\Phi(X) : H[Y|\Phi(X)] = H[Y|\Phi(X), \mathcal{E}]\} \tag{6}$$

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:

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

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

Remarks:

- $\Phi^*(X)$ is MIP.
- Optimal for OOD is $\hat{Y} = \mathbb{E}[Y|\Phi^*(X)]$.
- "Find $\Phi^*(X)$" $\rightarrow$ "Find MIP"

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

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

1 Background of OOD Generalization problem

2 Invariance-Based Optimization

3 Limitations

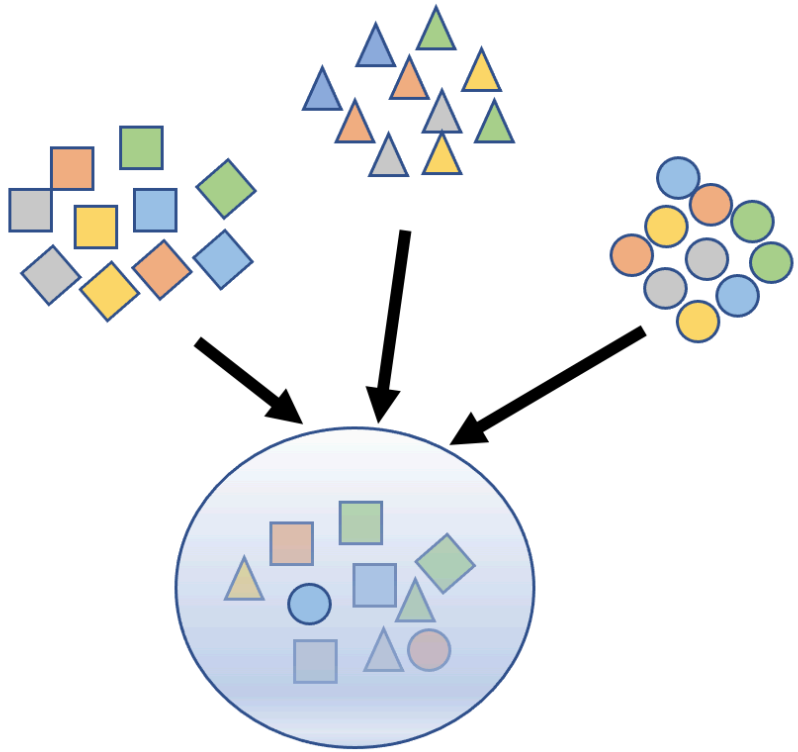
4 Heterogeneous Risk Minimization(HRM)

5 Kernelized Heterogeneous Risk Minimization(KerHRM)

6 Conclusion

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:

$$\arg\max_{\Phi \in \mathcal{I}_{\mathcal{E}}} I(Y; \Phi) \tag{8}$$

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. $\text{supp}(\mathcal{E}_{tr}) \subset \text{supp}(\mathcal{E})$
  2. $\mathcal{I}_{\mathcal{E}} \subset \mathcal{I}_{\mathcal{E}_{tr}}$
  3. Φ^*_{tr} NOT INVARIANT.

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

Example

	Class 0 (Cats)			Class 1 (Dogs)		
Index	X_1	X_2	X_3	X_1	X_2	X_3
e_1	Cats	Water	Irma	Dogs	Grass	Eric
e_2	Cats	Grass	Eric	Dogs	Water	Irma
e_3	Cats	Water	Eric	Dogs	Grass	Irma
e_4	Cats	Grass	Irma	Dogs	Water	Eric
e_5	Mixture: 90% data from e_1 and 10% data from e_2					
e_6	Mixture: 90% data from e_3 and 10% data from e_4					

表 1: A Toy Example for the difference between $\mathcal{I}_{\mathcal{E}}$ and $\mathcal{I}_{\mathcal{E}_{tr}}$.

- When $\text{supp}(\mathcal{E}) = \{e_1, e_2, e_3, e_4, e_5, e_6\}$, $\mathcal{I}_{\mathcal{E}} = \{\Phi | \Phi = \Phi(X_1)\}$.
- When $\text{supp}(\mathcal{E}_{tr}) = \{e_5, e_6\}$, $\mathcal{I}_{\mathcal{E}_{tr}} = \{\Phi | \Phi = \Phi(X_1, X_2)\}$.
- When e_5 and e_6 can be further divided into e_1, e_2 and e_3, e_4 respectively, $\mathcal{I}_{\mathcal{E}_{tr}}$ becomes $\mathcal{I}_{\mathcal{E}_{tr}} = \mathcal{I}_{\mathcal{E}} = \{\Phi(X_1)\}$.

What Kind of Environments is Needed?

- Recall the limitations:
  1. No training environments.
  2. Despite having environments, the quality of given environments.
- What kind of environments is needed?
  1. As heterogeneous as possible.
  2. Make $\mathcal{I}_{\mathcal{E}_{tr}}$ as close to $\mathcal{I}_{\mathcal{E}}$ as possible.
- How to **Generate Environments**?
  1. Randomly Split?
  2. ...

1 Background of OOD Generalization problem

2 Invariance-Based Optimization

3 Limitations

4 Heterogeneous Risk Minimization(HRM)

5 Kernelized Heterogeneous Risk Minimization(KerHRM)

6 Conclusion

HRM Problem

Assumption (Heterogeneity Assumption)

For random variable pair (X, Φ^*) and Φ^* satisfying the Invariance Assumption, using functional representation lemma[4], 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)

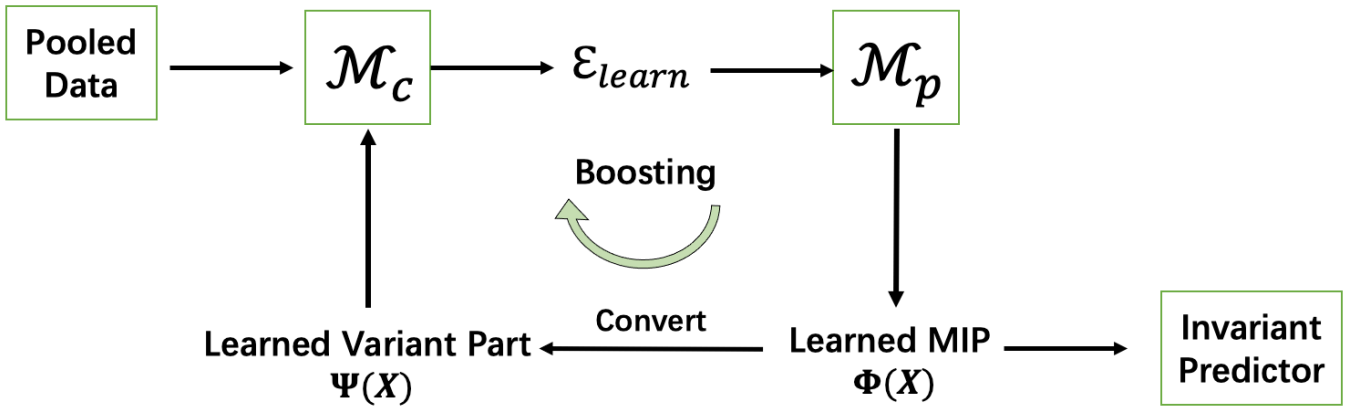
Given heterogeneous dataset $D = \{D^e\}_{e \in \mathrm{supp}(\mathcal{E}_{latent})}$ without environment labels, the task is to generate environments $\mathcal{E}_{learn}$ with minimal $|\mathcal{I}_{\mathcal{E}_{learn}}|$ and learn invariant model under learned $\mathcal{E}_{learn}$ with good OOD performance.

- 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.

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

The Whole Algorithm[5]

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

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

The Heterogeneity Identification Module $\mathcal{M}_c$

Recall that for $\mathcal{M}_c$,

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

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 $\mathcal{N}(f_{\Theta_j}(\Psi), \sigma^2)$:

$$h_j(\Psi, Y) = P_{\Theta_j}(Y|\Psi) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(Y - f_{\Theta_j}(\Psi))^2}{2\sigma^2}) \tag{9}$$

- 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:

$$\min_{Q\in\mathcal{Q}} D_{KL}(\hat{P}_N \| Q) \tag{10}$$

The Invariant Prediction Module $\mathcal{M}_p$

Recall that for $\mathcal{M}_p$,

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

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

- For invariant prediction, we adopt the regularizer[6] as:

$$\mathcal{L}_p(M \odot X, Y; \theta) = \mathbb{E}_{\mathcal{E}_{tr}}[\mathcal{L}^e] + \lambda \text{trace}(\text{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e)) \tag{11}$$

  - 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:

$$\mathcal{L}^e(\theta, \mu) = \mathbb{E}_{P^e}\mathbb{E}_M\left[\ell(M \odot X^e, Y^e; \theta) + \alpha\|M\|_0\right] \tag{12}$$

  - $\|M\|_0$ controls the number of selected features.
  - Conduct continuous optimization as [7].

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

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

The Mutual Promotion

- Insight: We should only use Ψ^* for Heterogeneity Identification.

Assumption (Heterogeneity Assumption from Information Theory)

Assume the pooled training data is made up of heterogeneous data sources: $P_{tr} = \sum_{e \in \mathrm{supp}(\mathcal{E}_{tr})} w_e P^e$. For any $e_i, e_j \in \mathcal{E}_{tr}, e_i \neq e_j$, we assume

$$I^c_{i,j}(Y; \Phi^* | \Psi^*) \geq \max(I_i(Y; \Phi^* | \Psi^*), I_j(Y; \Phi^* | \Psi^*)) \tag{13}$$

where Φ^* is invariant feature and Ψ^* the variant. I_i represents mutual information in P^{e_i} and $I^c_{i,j}$ represents the cross mutual information between P^{e_i} and P^{e_j} takes the form of $I^c_{i,j}(Y; \Phi | \Psi) = H^c_{i,j}[Y|\Psi] - H^c_{i,j}[Y|\Phi, \Psi]$ and $H^c_{i,j}[Y] = -\int p^{e_i}(y) \log p^{e_j}(y) dy$.

- The mutual information $I_i(Y; \Phi^*) = H_i[Y] - H_i[Y|\Phi^*]$ can be viewed as the error reduction if we use Φ^* to predict Y rather than predict by nothing.
- The cross mutual information $I^c_{i,j}(Y; \Phi^*)$ can be viewed as the error reduction if we use the predictor learned on Φ^* in environment e_j to predict in environment e_i, rather than predict by nothing.

Theorem (Why using only Ψ?)

For $e_i, e_j \in \mathrm{supp}(\mathcal{E}_{tr})$, assume that $X = [\Phi^*, \Psi^*]^T$ satisfying Invariance and Heterogeneity Assumption, where Φ^* is invariant and Ψ^* variant. Then we have $\mathrm{D_{KL}}(P^{e_i}(Y|X) \| P^{e_j}(Y|X)) \leq \mathrm{D_{KL}}(P^{e_i}(Y|\Psi^*) \| P^{e_j}(Y|\Psi^*))$

Experiment Results

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]):

$$\min_{\Phi} \max_{u} \sum_{e \in \mathcal{E}} \frac{1}{N_e} \sum_i u_i(e) \ell(w \odot \Phi(x_i), y_i) + \sum_{e \in \mathcal{E}} \lambda \| \nabla_{w|w=1.0} \frac{1}{N_e} \sum_i u_i(e) \ell(w \odot \Phi(x_i), y_i) \|_2 \tag{14}$$

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

$$\min_{\Phi} \sum_{e \in \mathcal{E}_{tr}} \mathcal{L}^e + \lambda \| \nabla_{w|w=1.0} \mathcal{L}^e(w \odot \Phi) \|^2 \tag{15}$$

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 $V_b \subset \Psi^*$:

$$\hat{P}(x) = \prod_{v_i \in V_b} |r|^{-5*|f(\phi^*) - sign(r)*v_i|} \tag{16}$$

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.

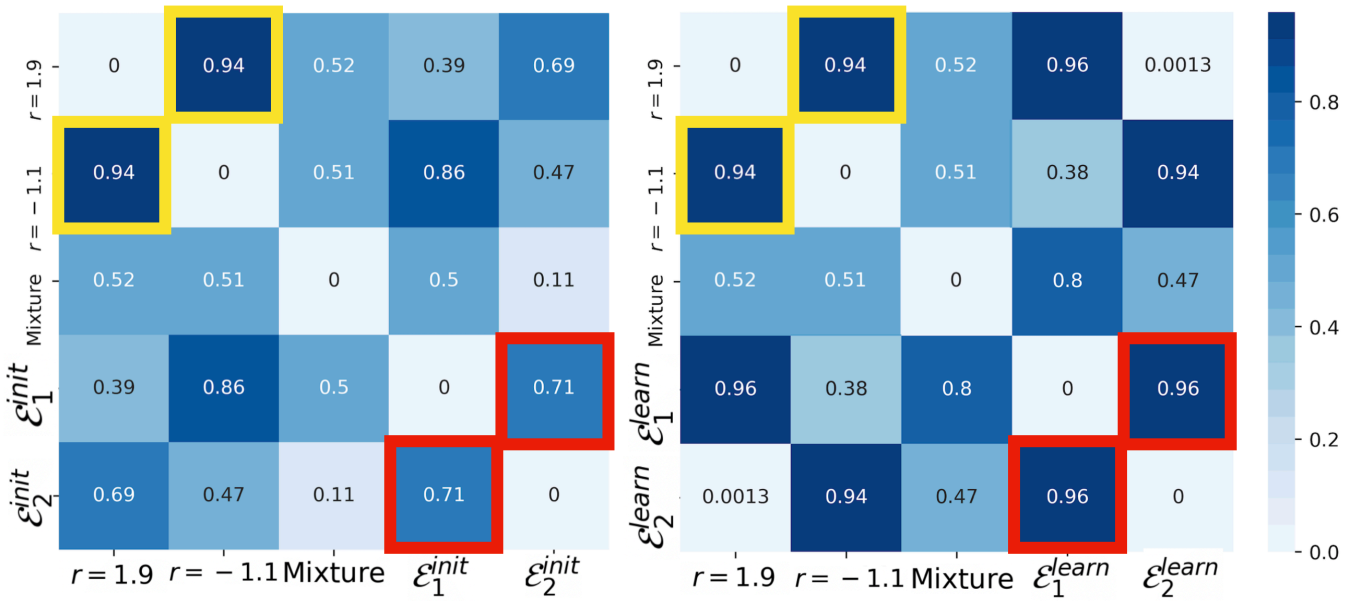
Selection Bias Results

表 2: 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[8] as follows:



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

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

Anti-Causal Effect

- Setting: $X = [\Phi^*, \Psi^*]^T \in \mathbb{R}^d$ and firstly sample Φ^* from $\sum_{i=1}^{k} z_k \mathcal{N}(\mu_i, I)$ and $Y = \theta_\phi^T \Phi^* + \beta \Phi_1 \Phi_2 \Phi_3 + \mathcal{N}(0, 0.3)$. Then the spurious correlations between Ψ^* and Y are generated by anti-causal effect as

$$\Psi^* = \theta_\psi Y + \mathcal{N}(0, \sigma(\mu_i)^2) \tag{17}$$

  - The larger the $\sigma(\mu_i)$ is, the weaker correlation between Ψ^* and Y.

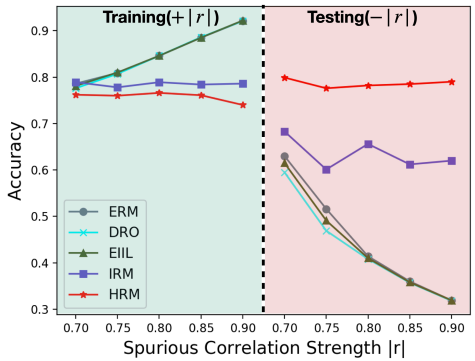
表 3: Prediction errors of the anti-causal effect experiment. We design two settings with different dimensions of Φ^* and Ψ^* as n_ϕ and n_ψ respectively. The results are averaged over 10 runs.

Scenario 1: $n_\phi = 9,\ n_\psi = 1$										
e	Training environments			Testing environments						
Methods	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}
ERM	0.290	0.308	0.376	0.419	0.478	0.538	0.596	0.626	0.640	0.689
DRO	0.289	0.310	0.388	0.428	0.517	0.610	0.627	0.669	0.679	0.739
EIIL	**0.075**	**0.128**	0.349	0.485	0.795	1.162	1.286	1.527	1.558	1.884
IRM(with $\mathcal{E}_{tr}$ label)	0.306	0.312	0.325	0.328	0.343	0.358	0.365	0.374	0.377	0.392
HRM^s	1.060	1.085	1.112	1.130	1.207	1.280	1.325	1.340	1.371	1.430
HRM	0.317	0.314	**0.322**	**0.318**	**0.321**	**0.317**	**0.315**	**0.315**	**0.316**	**0.320**
Scenario 2: $n_\phi = 5,\ n_\psi = 5$										
e	Training environments			Testing environments						
Methods	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}
ERM	0.238	0.286	0.433	0.512	0.629	0.727	0.818	0.860	0.895	0.980
DRO	0.237	0.294	0.452	0.529	0.651	0.778	0.859	0.911	0.950	1.028
EIIL	**0.043**	**0.145**	0.521	0.828	1.237	1.971	2.523	2.514	2.506	3.512
IRM(with $\mathcal{E}_{tr}$ label)	0.287	0.293	0.329	0.345	0.382	0.420	0.444	0.461	0.478	0.504
HRM^s	0.455	0.463	0.479	0.478	0.495	0.508	0.513	0.519	0.525	0.533
HRM	0.316	0.315	**0.315**	**0.330**	**0.320**	**0.317**	**0.326**	**0.330**	**0.333**	**0.335**

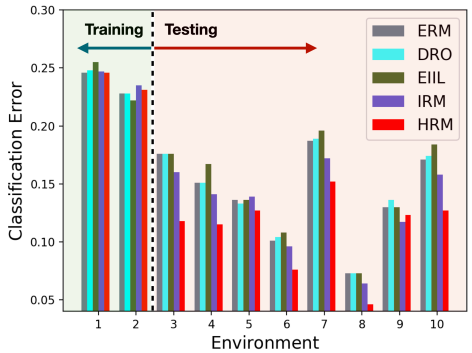
Real-World Datasets

Datasets:

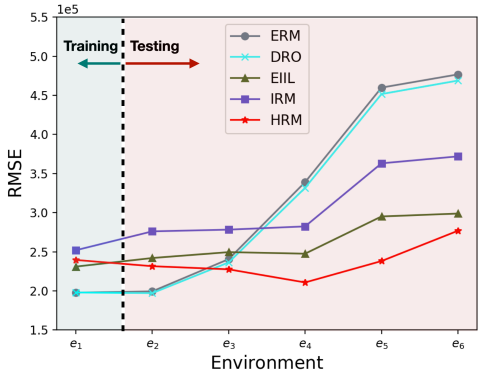
- Car Insurance[9]
- People Income Prediction[10]
- House Price Prediction[11]



(a) Training and testing accuracy for the car insurance prediction. Left sub-figure shows the training results for 5 settings and the right shows their corresponding testing results.



(b) Mis-Classification Rate for the income prediction.



(c) Prediction error for the house price prediction. RMSE refers to the Root Mean Square Error.

图 1: Results of real-world datasets, including training and testing performance for five methods.

[9] https://www.kaggle.com/anmolkumar/health-insurance- cross-sell-prediction

[10] Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[11] https://www.kaggle.com/c/house-prices-advanced- regression- techniques/data

1 Background of OOD Generalization problem

2 Invariance-Based Optimization

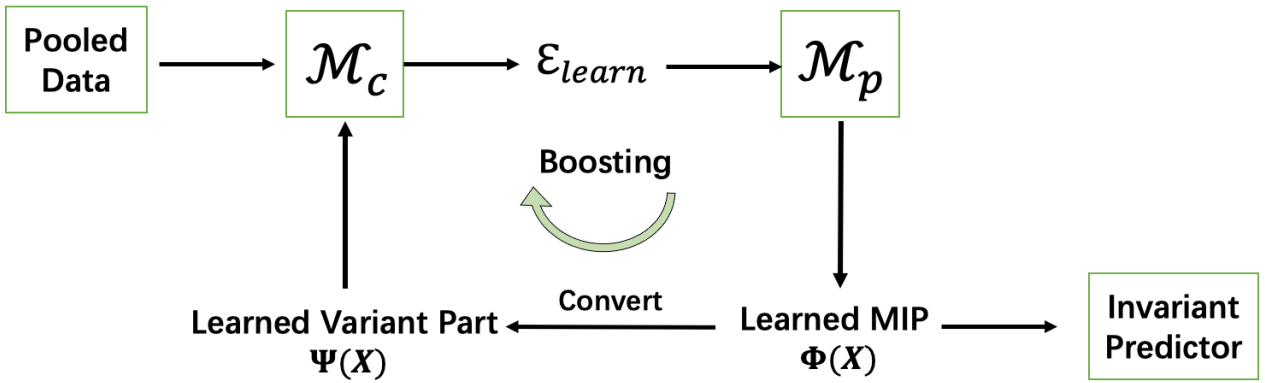
3 Limitations

4 Heterogeneous Risk Minimization(HRM)

5 Kernelized Heterogeneous Risk Minimization(KerHRM)

6 Conclusion

Limitations of HRM



- $\Phi(X) = M \odot X$ and $\Psi(X) = (1 - M) \odot X$
- HRM can only deal with raw feature data, when features are mixed (e.g. $X = H[S, V]^T$) such feature selection procedure will fail.

The Goal of KerHRM: Extend HRM to Complicated Data

Preliminaries

- **Kernel function**: A kernel is a function κ that for all $x, z \in \mathcal{X}$ satisfies

$$\kappa(x, z) =< \phi(x), \phi(z) > \tag{18}$$

where ϕ is a mapping from $\mathcal{X}$ to an inner product feature space $\mathcal{H}$: $\phi : x \mapsto \phi(x) \in \mathcal{H}$.

  - Judgement: Gram 矩阵 ($K_{i,j} = \kappa(x_i, x_j)$)
  - Linear kernel: $\kappa_{lin}(x, y) = x^T y$
  - Gaussian-RBF kernel: $\kappa(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma^2})$

$$\phi(x_i) = \exp(-x_i^2)[1, \sqrt{\frac{2}{1!}}x_i, \sqrt{\frac{2^2}{2!}}x_i^2, \sqrt{\frac{2^3}{3!}}x_i^3 \ldots] \tag{19}$$

- **Representer Theorem**: Consider linear model differentiable with losses ℓ_i and L_2 regularizer:

$$\arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \ell_i(w^T x_i) + \frac{\lambda}{2}\|w\|^2 \tag{20}$$

Any solution w^* can be written as a **linear combination of features** x_i:

$$w^* = \frac{1}{\lambda}\sum_{i=1}^{n} \frac{\partial \ell_i}{\partial w}(w^T x_i)x_i = \sum_{i=1}^{n} z_i x_i = X^T z \tag{21}$$

Preliminaries: Ridge Regression from Kernel Perspective

- Ridge Regression Objective:

$$\min_{w\in\mathbb{R}^d} \frac{1}{2}\|Xw - Y\|^2 + \frac{\lambda}{2}\|w\|^2 \tag{22}$$

The minimum is given by $w = (X^TX + \lambda I)^{-1}_{d\times d}X^TY = X^T(XX^T + \lambda I)^{-1}_{n\times n}Y$

- For new data $\tilde{X} \in \mathbb{R}^{t\times d}$, the prediction is:

$$\begin{aligned}\tilde{Y} &= \tilde{X}w \in \mathbb{R}^t \\ &= \tilde{X}X^T(XX^T + \lambda I)Y \\ &= \tilde{K}_{t\times n}(K + \lambda I)^{-1}_{n\times n}Y\end{aligned} \tag{23}$$

  - $\phi(x), \phi(z)$ is hidden in $\kappa(x, z)$

- Another formulation: according to the representer theorem, we have $w = \sum_{i=1}^n z_i x_i$, then the prediction becomes:

$$\tilde{Y}(x) = \sum_{i=1}^n z_i \kappa(x, x_i) \tag{24}$$

- Kernel can measure the similarity between data.

Preliminaries: Neural Tangent Kernel

- From MLP to Linear:

$$f(x, \mathbf{w}) \approx f(x, w_0) + \nabla_w f(x, w_0)^T (\mathbf{w} - w_0) \tag{25}$$

  - lazy training makes such Taylor expansion not approximate too much
  - $\nabla_w f(x, w_0)^T \in \mathbb{R}^{n \times p}$

- Neural Tangent Kernel(κ_{NTK}):

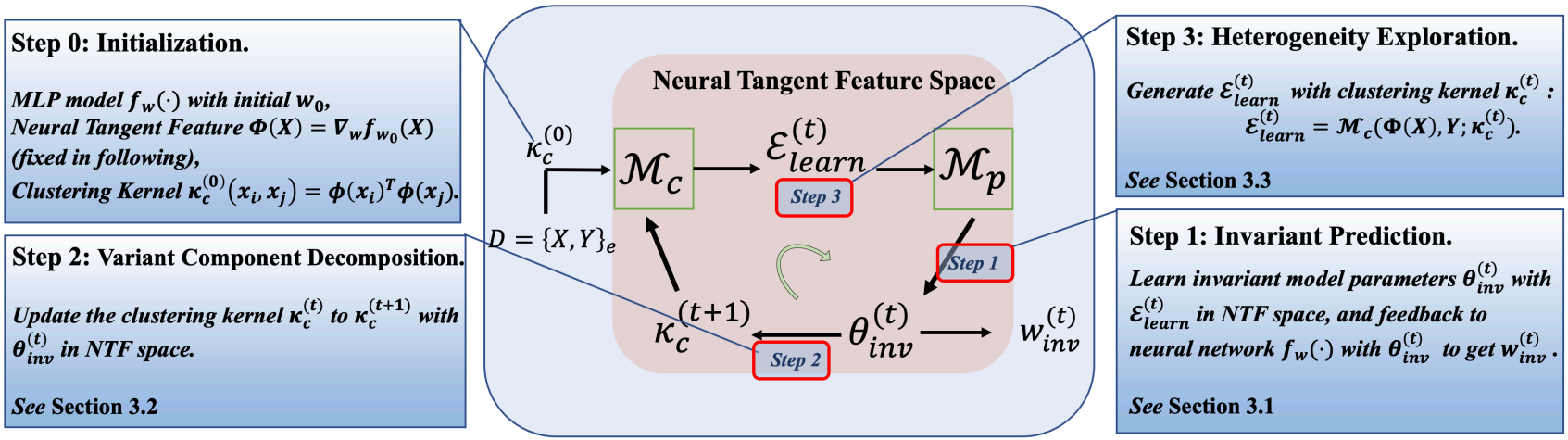
$$\kappa_{NTK}(x_i, x_j) = \mathbb{E}_{w_0} [< \nabla_w f(x_i, w_0), \nabla_w f(x_j, w_0) >] \tag{26}$$

  - $\mathbb{E}_{w_0}$ is not important, since $w - w_0$ is quite tiny due to lazy training
  - X is mapped to $\nabla_w f(X, w_0)^T \in \mathbb{R}^{n \times p}$. Define $\Phi(X)$ as:

$$\Phi(X)^T = \nabla_w f(X, w_0)^T \in \mathbb{R}^{n \times p} \tag{27}$$

Idea: Convert the input space to Neural Tangent Feature Space.

Kernelized Heterogeneous Risk Minimization(KerHRM[12])



- Step 0:

$$f_w(X) \approx f_{w_0}(X) + \nabla_w f_{w_0}(X)^T (w - w_0) \tag{28}$$
$$= f_{w_0}(X) + \Phi(X)^T (w - w_0) \tag{29}$$
$$\approx f_{w_0}(X) + USV^T (w - w_0) \tag{30}$$
$$= f_{w_0}(X) + \Psi(X)\left(V^T (w - w_0)\right) = f_{w_0}(X) + \Psi(X)\theta \tag{31}$$

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.

[12] 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 $\Psi(X)$[13]:

$$\theta_{inv} = \arg\min_{\theta} \sum_{e \in \mathcal{E}_{learn}} \mathcal{L}^e(\theta; \Psi, Y) + \alpha \text{Var}_{\mathcal{E}_{learn}}(\nabla_\theta \mathcal{L}^e) \tag{32}$$

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:

$$\kappa_c^{(0)}(x_i, x_j) = \phi(x_i)^T \phi(x_j) = \langle U_i S, U_j S \rangle \tag{33}$$

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

$$\Psi_V^{(t+1)}(x_i) \leftarrow U_i S - \left\langle U_i S, \theta_{inv}^{(t)} \right\rangle \theta_{inv}^{(t)} / \|\theta_{inv}^{(t)}\|^2 \tag{34}$$

  - Obtain a new kernel for clustering:

$$\kappa_c^{(t+1)}(x_i, x_j) = \Psi_V^{(t+1)}(x_i)^T \Psi_V^{(t+1)}(x_j) \tag{35}$$

[13] Here 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:

$$h_j(\Psi_V(X), Y) = P_{Theta_j}(Y|\Psi_V(X)) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(Y - f(\Theta_j; \Psi_V(X)))^2/2]\sigma^2) \tag{36}$$

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

$$\mathcal{Q} = \{Q : Q = \sum_{k\in[K]} q_j h_j\} \tag{37}$$

and gives the objective function:

$$\min_{Q\in\mathcal{Q}} D_{KL}(\hat{P}_N \| Q) \Leftrightarrow \min_{\Theta,\mathbf{q}} \left\{ \mathcal{L}_c = -\frac{1}{N} \sum_{i\in[N]} \log \left[\sum_{j\in[K]} q_j h_j(\psi_V(x_i), y_i) \right] \right\} \tag{38}$$

Experiment Results

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]):

$$\min_{\Phi} \max_{u} \sum_{e \in \mathcal{E}} \frac{1}{N_e} \sum_i u_i(e) \ell(w \odot \Phi(x_i), y_i) + \sum_{e \in \mathcal{E}} \lambda \| \nabla_{w|w=1.0} \frac{1}{N_e} \sum_i u_i(e) \ell(w \odot \Phi(x_i), y_i) \|_2 \tag{39}$$

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

$$\min_{\Phi} \sum_{e \in \mathcal{E}_{tr}} \mathcal{L}^e + \lambda \| \nabla_{w|w=1.0} \mathcal{L}^e(w \odot \Phi) \|^2 \tag{40}$$

Simulation: Classification with Spurious Correlation

- Data Generation:

$$S|Y \sim \mathcal{N}(Y\mathbf{1}, \sigma_s^2 I_d), \ \ V|A \sim \mathcal{N}(A\mathbf{1}, \sigma_v^2 I_d) \tag{41}$$

where the label $Y \in \{+1, -1\}$ and the spurious attribute $A \in \{+1, -1\}$. Each environment is characterized by its bias rate $r \in (0, 1]$ (for $100 * r\%$ data, $A = Y$, for others, $A = -Y$).

$$X = H[S, V]^T \in \mathbb{R}^{2d} \tag{42}$$

where $H \in \mathbb{R}^{2d \times 2d}$ is a random orthogonal matrix.

- Settings: for training, $r_1 = 0.9, r_2 = r$; for testing, $r_3 = 0.1$.

Results

Table 1: Results in classification simulation experiments of different methods with varying bias rate r_2, and scrambled matrix H, and each result is averaged over ten times runs.

r_2	$r_2 = 0.70$		$r_2 = 0.75$		$r_2 = 0.80$	
Methods	Train_Acc	Test_Acc	Train_Acc	Test_Acc	Train_Acc	Test_Acc
ERM	0.850	0.400	0.862	0.325	0.875	0.254
DRO	0.857	0.473	0.870	0.432	0.883	0.395
EIIL	**0.927**	0.523	**0.925**	0.470	**0.946**	0.463
HRM	0.836	0.543	0.832	0.519	0.852	0.488
IRM(with $\mathcal{E}_{tr}$ label)	0.836	0.606	0.853	0.544	0.877	0.401
KerHIL^s	0.764	0.671	0.782	0.632	0.663	0.619
KerHIL	0.759	**0.724**	0.760	**0.686**	0.741	**0.693**

Simulation: Colored MNIST

Data Generation[14]:

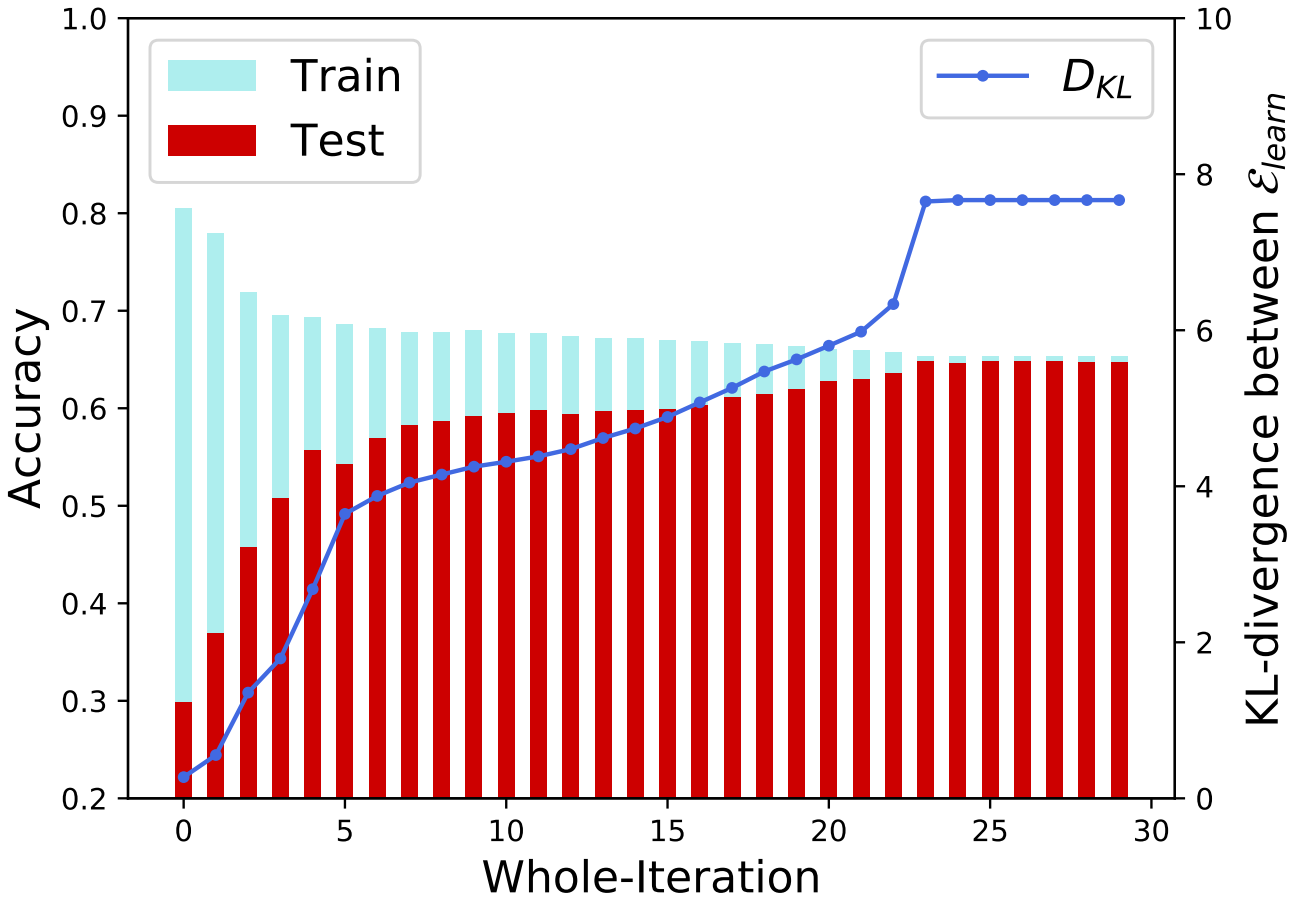
- Firstly, a binary label Y is assigned to each images according to its digits: $Y = 0$ for digits 0∼4 and $Y = 1$ for digits 5∼9.
- Secondly, we sample the color id C by flipping Y with probability e and therefore forms environments, where $e = 0.1$ for the first training environment, $e = 0.2$ for the second training environments and $e = 0.9$ for the testing environment.
- Thirdly, we induce noisy labels by randomly flipping the label Y with probability 0.2.

Settings:

- For training, $e_1 = 0.1, e_2 = 0.2$.
- For testing, $e_3 = 0.9$.
- The correlation between Color and Label is inverse between training and testing.

[14] Martin Arjovsky *et al.* Invariant Risk Minimization

Surprising Results



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

1 Background of OOD Generalization problem

2 Invariance-Based Optimization

3 Limitations

4 Heterogeneous Risk Minimization(HRM)

5 Kernelized Heterogeneous Risk Minimization(KerHRM)

6 Conclusion

Conclusion

As for HRMs, there exist some drawbacks:

- The convergence guarantee of the whole framework, especially the frontend, still remains ambiguous.
- Extend to more complicated data?

As for the OOD Generalization problem, there exist some open problems

- How to formulate the OOD Generalization problem? How to justify its learnability?
- Environment complexity of methods for OOD Generalization.
- Real datasets to evaluate the effect of methods for OOD Generalization?
- Incorporate pre-trained models to deal with complicated data?

Some other materials

- Anual Progress Report on Out-of-Distribution Generalization[15]
- Stable Learning and its Causal Implication[16]

[15] http://pengcui.thumedialab.com/papers/OOD_APR_valse2021.pdf

[16] http://pengcui.thumedialab.com/papers/Stable%20Learning-tutorial-valse2021.pdf

Contact

Jiashuo Liu

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