Heterogeneous Risk Minimization

Jiashuo Liu [1] **Zheyuan Hu** [1] **Peng Cui** [1] **Bo Li** [2] **Zheyan Shen** [1]

Abstract

Machine learning algorithms with empirical risk minimization usually suffer from poor generalization performance due to the greedy exploitation of correlations among the training data, which are not stable under distributional shifts. Recently, some invariant learning methods for out-of-distribution (OOD) generalization have been proposed by leveraging multiple training environments to find invariant relationships. However, modern datasets are frequently assembled by merging data from multiple sources without explicit source labels. The resultant unobserved heterogeneity renders many invariant learning methods inapplicable. In this paper, we propose Heterogeneous Risk Minimization (HRM) framework to achieve joint learning of latent heterogeneity among the data and invariant relationship, which leads to stable prediction despite distributional shifts. We theoretically characterize the roles of the environment labels in invariant learning and justify our newly proposed HRM framework. Extensive experimental results validate the effectiveness of our HRM framework.

1. Introduction

The effectiveness of machine learning algorithms with empirical risk minimization (ERM) relies on the assumption that the testing and training data are identically drawn from the same distribution, which is known as the IID hypothesis. However, distributional shifts between testing and training data are usually inevitable due to data selection biases or unobserved confounders that widely exist in real data. Under such circumstances, machine learning algorithms with ERM usually suffer from poor generalization performance due to the greedy exploitation of correlations among the training data, which are not stable under distributional shifts. How to guarantee a machine learning algorithm with out-of-distribution (OOD) generalization ability and stable performances under distributional shifts is of paramount significance, especially in high-stake applications such as medical diagnosis, criminal justice, and financial analysis etc (Kukar, 2003; Berk et al., 2018; Rudin & Ustun, 2018).

There are mainly two branches of methods proposed to solve the OOD generalization problem, namely distributionally robust optimization (DRO) (Esfahani & Kuhn, 2018; Duchi & Namkoong, 2018; Sinha et al., 2018; Sagawa et al., 2019) and invariant learning (Arjovsky et al., 2019; Koyama & Yamaguchi, 2020; Chang et al., 2020). DRO methods aim to optimize the worst-performance over a distribution set to ensure their OOD generalization performances. While DRO is a powerful family of methods, it is often argued for its over-pessimism problem when the distribution set is large (Hu et al., 2018; Frogner et al., 2019). From another perspective, invariant learning methods propose to exploit the causally invariant correlations(rather than varying spurious correlations) across multiple training environments, resulting in out-of-distribution (OOD) optimal predictors. However, the effectiveness of such methods relies heavily on the quality of training environments, and the intrinsic role of environments in invariant learning remains vague in theory. More importantly, modern big data are frequently assembled by merging data from multiple sources without explicit source labels. The resultant unobserved heterogeneity renders these invariant learning methods inapplicable.

In this paper, we propose Heterogeneous Risk Minimization (HRM), an optimization framework to achieve joint learning of the latent heterogeneity among the data and the invariant predictor, which leads to better generalization ability despite distributional shifts. More specifically, we theoretically characterize the roles of the environment labels in invariant learning, which motivates us to design two modules in the framework corresponding to heterogeneity identification and invariant learning respectively. We provide theoretical justification on the mutual promotion of these two modules, which resonates the joint optimization process in a reciprocal way. Extensive experiments in both synthetic and real-world experiments datasets demonstrate

[1] Department of Computer Science and Technology, Tsinghua University, Beijing, China; Email: {liujiashuo77, zyhu2001}@gmail.com, cuip@tsinghua.edu.cn, shenzy17@mails.tsinghua.edu.cn. [2] School of Economics and Management, Tsinghua University, Beijing, China; Email: libo@sem.tsinghua.edu.cn. Correspondence to: Peng Cui <cuip@tsinghua.edu.cn>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

the superiority of HRM in terms of average performance, stability performance as well as worst-case performance under different settings of distributional shifts. We summarize our contributions as following:

1. We propose the novel HRM framework for OOD generalization without environment labels, in which heterogeneity identification and invariant prediction are jointly optimized.

2. We theoretically characterize the role of environments in invariant learning from the perspective of heterogeneity, based on which we propose a novel clustering method for heterogeneity identification from heterogeneous data.

3. We theoretically justify the mutual promotion relationship between heterogeneity identification and invariant learning, resonating the joint optimization process in HRM.

2. Problem Formulation

2.1. OOD and Maximal Invariant Predictor

Following (Arjovsky et al., 2019; Chang et al., 2020), we consider a dataset $D = \{D^e\}_{e\in\text{supp}(\mathcal{E}_{tr})}$, which is a mixture of data $D^e = \{(x_i^e, y_i^e)\}_{i=1}^{n_e}$ collected from multiple training environments $e \in \text{supp}(\mathcal{E}_{tr})$, $x_i^e \in \mathcal{X}$ and $y_i^e \in \mathcal{Y}$ are the i-th data and label from environment e respectively and n_e is number of samples in environment e. Environment labels are unavailable as in most real applications. $\mathcal{E}_{tr}$ is a random variable on indices of training environments and P^e is the distribution of data and label in environment e.

The goal of this work is to find a predictor $f(\cdot) : \mathcal{X} \to \mathcal{Y}$ with good out-of-distribution generalization performance, which can be formalized as:

$$\arg\min_f \max_{e\in\text{supp}(\mathcal{E})} \mathcal{L}(f|e) \tag{1}$$

where $\mathcal{L}(f|e) = \mathbb{E}[l(f(X), Y)|e] = \mathbb{E}^e[l(f(X^e), Y^e)]$ is the risk of predictor f on environment e, and $l(\cdot,\cdot) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ is the loss function. $\mathcal{E}$ is the random variable on indices of all possible environments such that $\text{supp}(\mathcal{E}) \supset \text{supp}(\mathcal{E}_{tr})$. Usually, for all $e \in \text{supp}(\mathcal{E}) \setminus \text{supp}(\mathcal{E}_{tr})$, the data and label distribution $P^e(X, Y)$ can be quite different from that of training environments $\mathcal{E}_{tr}$. Therefore, the problem in Equation 1 is referred to as Out-of-Distribution (OOD) Generalization problem (Arjovsky et al., 2019).

Without any prior knowledge or structural assumptions, it is impossible to figure out the OOD generalization problem, since one cannot characterize the unseen latent environments in $\text{supp}(\mathcal{E})$. A commonly used assumption in invariant learning literature (Rojas-Carulla et al., 2015; Gong et al., 2016; Arjovsky et al., 2019; Kuang et al., 2020; Chang et al., 2020) is as follow:

Assumption 2.1. *There exists random variable $\Phi^*(X)$ such that the following properties hold:*

a. Invariance property*: for all $e, e' \in \text{supp}(\mathcal{E})$, we have $P^e(Y|\Phi^*(X)) = P^{e'}(Y|\Phi^*(X))$ holds.*

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

This assumption indicates invariance and sufficiency for predicting the target Y using Φ^*, which is known as invariant covariates or representations with stable relationships with Y across different environments $e \in \mathcal{E}$.

In order to acquire the invariant predictor $\Phi^*(X)$, a branch of work to find maximal invariant predictor (Chang et al., 2020; Koyama & Yamaguchi, 2020) has been proposed, where the invariance set and the corresponding maximal invariant predictor are defined as:

Definition 2.1. *The invariance set $\mathcal{I}$ with respect to $\mathcal{E}$ is defined as:*

$$\begin{aligned}\mathcal{I}_{\mathcal{E}} &= \{\Phi(X) : Y \perp \mathcal{E}|\Phi(X)\} \\ &= \{\Phi(X) : H[Y|\Phi(X)] = H[Y|\Phi(X), \mathcal{E}]\}\end{aligned} \tag{2}$$

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:

$$S = \arg\max_{\Phi\in\mathcal{I}_{\mathcal{E}}} I(Y; \Phi) \tag{3}$$

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

Here we prove that the MIP S can can guarantee OOD optimality, as indicated in Theorem 2.1. The formal statement of Theorem 2.1 as well as its proof can be found in appendix.

Theorem 2.1. *(Informal) For predictor $\Phi^*(X)$ satisfying Assumption 2.1, Φ^* is the maximal invariant predictor with respect to $\mathcal{E}$ and the solution to OOD problem in equation 1 is $\mathbb{E}_Y[Y|\Phi^*] = \arg\min_f \sup_{e\in\text{supp}(\mathcal{E})} \mathbb{E}[\mathcal{L}(f)|e]$.*

Recently, some works suppose the availability of data from multiple environments with environment labels, wherein they can find MIP (Chang et al., 2020; Koyama & Yamaguchi, 2020). However, they rely on the underlying assumption that the invariance set $\mathcal{I}_{\mathcal{E}_{tr}}$ of $\mathcal{E}_{tr}$ is exactly the invariance set $\mathcal{I}_{\mathcal{E}}$ of all possible unseen environments $\mathcal{E}$, which cannot be guaranteed as shown in Theorem 2.2.

Theorem 2.2. $\mathcal{I}_{\mathcal{E}} \subseteq \mathcal{I}_{\mathcal{E}_{tr}}$

As shown in Theorem 2.2 that $\mathcal{I}_{\mathcal{E}} \subseteq \mathcal{I}_{\mathcal{E}_{tr}}$, the learned predictor is only invariant to such limited environments $\mathcal{E}_{tr}$ but is not guaranteed to be invariant with respect to all possible environments $\mathcal{E}$.

Here we give a toy example in Table 1 to illustrate this. We consider a binary classification between cats and dogs, where each photo contains 3 features, animal feature $X_1 \in \{\text{cat}, \text{dog}\}$, a background feature $X_2 \in \{\text{on grass}, \text{in water}\}$ and the photographer's signature feature $X_3 \in \{\text{Irma}, \text{Eric}\}$. Assume all possible testing

	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					

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

environments $\text{supp}(\mathcal{E}) = \{e_1, e_2, e_3, e_4, e_5, e_6\}$ and the train environment $\text{supp}(\mathcal{E}_{tr}) = \{e_5, e_6\}$, then $\mathcal{I}_{\mathcal{E}} = \{\Phi | \Phi = \Phi(X_1)\}$ while $\mathcal{I}_{\mathcal{E}_{tr}} = \{\Phi | \Phi = \Phi(X_1, X_2)\}$. The reason is that e_5, e_6 only tell us X_3 cannot be included in the invariance set but cannot exclude X_2. But if e_5 and e_6 can be further divided into e_1, e_2 and e_3, e_4 respectively, the invariance set $\mathcal{I}_{\mathcal{E}_{tr}}$ becomes $\mathcal{I}_{\mathcal{E}_{tr}} = \mathcal{I}_{\mathcal{E}} = \{\Phi(X_1)\}$.

This example shows that the manually labeled environments may not be sufficient to achieve MIP, not to mention the cases where environment labels are not available. This limitation necessitates the study on how to exploit the latent intrinsic heterogeneity in training data (like e_5 and e_6 in the above example) to form more refined environments for OOD generalization. The environments need to be subtly uncovered, in the sense of OOD generalization problem, as indicated by Theorem 2.3, not all environments are helpful to tighten the invariance set.

Theorem 2.3. *Given set of environments $\text{supp}(\hat{\mathcal{E}})$, denote the corresponding invariance set $\mathcal{I}_{\hat{\mathcal{E}}}$ and the corresponding maximal invariant predictor $\hat{\Phi}$. For one newly-added environment e_{new} with distribution $P^{new}(X, Y)$, if $P^{new}(Y|\hat{\Phi}) = P^e(Y|\hat{\Phi})$ for $e \in \text{supp}(\hat{\mathcal{E}})$, the invariance set constrained by $\text{supp}(\hat{\mathcal{E}}) \cup \{e_{new}\}$ is equal to $\mathcal{I}_{\hat{\mathcal{E}}}$.*

2.2. Problem of Heterogeneous Risk Minimization

Besides Assumption 2.1, we make another assumption on the existence of heterogeneity in training data as:

Assumption 2.2. Heterogeneity Assumption.
For random variable pair (X, Φ^*) and Φ^* satisfying Assumption 2.1, using functional representation lemma (El Gamal & Kim, 2011), there exists random variable Ψ^* such that $X = X(\Phi^*, \Psi^*)$, then we assume $P^e(Y|\Psi^*)$ can arbitrary change across environments $e \in \text{supp}(\mathcal{E})$.

The heterogeneity among provided environments can be evaluated by the compactness of the corresponding invariance set as $|\mathcal{I}_{\mathcal{E}}|$. Specifically, smaller $|\mathcal{I}_{\mathcal{E}}|$ leads to higher heterogeneity, since more variant features can be excluded. Based on the assumption, we come up with the problem of heterogeneity exploitation for OOD generalization.

Problem 1. *Heterogeneous Risk Minimization. Given heterogeneous dataset $D = \{D^e\}_{e \in \text{supp}(\mathcal{E}_{latent})}$ without environment labels, the task is to generate environments $\mathcal{E}_{tr}$ with minimal $|\mathcal{I}_{\mathcal{E}_{tr}}|$ and learn invariant model under learned $\mathcal{E}_{tr}$ with good OOD performance.*

Theorem 2.3 together with Assumption 2.2 indicate that, to better constrain $\mathcal{I}_{\mathcal{E}_{tr}}$, the effective way is to generate environments with varying $P(Y|\Psi^*(X))$ that can exclude variant features from $\mathcal{I}_{\mathcal{E}_{tr}}$. Under this problem setting, we encounter the circular dependency: first we need variant Ψ^* to generate heterogeneous environments $\mathcal{E}_{tr}$; then we need $\mathcal{E}_{tr}$ to learned invariant Φ^* as well as variant Ψ^*. Furthermore, there exists positive feedback between these two steps. When acquiring $\mathcal{E}_{tr}$ with tighter $\mathcal{I}_{\mathcal{E}_{tr}}$, more invariant predictor $\Phi(X)$ (i.e. a better approximation of MIP) can be found, which will further bring a clearer picture of variant parts, and therefore promote the generation of $\mathcal{E}_{tr}$. With this notion, we propose our framework for Heterogeneous Risk Minimization (HRM) which leverages the mutual promotion between the two steps and conduct joint optimization.

3. Method

In this work, we temporarily focus on a simple but general setting, where $X = [\Phi^*, \Psi^*]^T \in \mathbb{R}^d$ in raw feature level and Φ^*, Ψ^* satisfy Assumption 2.1. Under this setting, Our Heterogeneous Risk Minimization (HRM) framework contains two interactive parts, the frontend $\mathcal{M}_c$ for heterogeneity identification and the backend $\mathcal{M}_p$ for invariant prediction. The general framework is shown in Figure 1. Given the pooled heterogeneous data, it starts with the heterogeneity identification module $\mathcal{M}_c$ leveraging the learned variant representation $\Psi(X)$ to generate heterogeneous environments $\mathcal{E}_{learn}$. Then the learned environments are used by OOD prediction module $\mathcal{M}_p$ to learn the MIP $\Phi(X)$ as well as the invariant prediction model $f(\Phi(X))$. After that, we derive the variant $\Psi(X)$ to further boost the module $\mathcal{M}_c$, which is supported by Theorem 2.3. As for the 'convert' step, under our setting, we adopt feature selection in this work, through which more variant feature Ψ can be attained when more invariant feature Φ is learned. Specifically, the invariant predictor $\Phi(X)$ is generated as $\Phi(X) = M \odot X$, and the variant part $\Psi(X) = (1 - M) \odot X$ correspondingly, where $M \in \{0, 1\}^d$ is the binary invariant feature selection mask. For instance, for Table 1, $X = [X_1, X_2, X_3]$, the

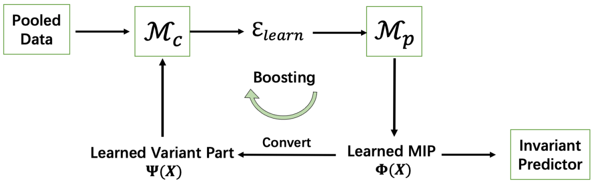


Figure 1. The framework of HRM.

ground truth binary mask is $M = [1, 0, 0]$. In this way, the better Φ is learned, the better Ψ can be obtained. Note that we use the soft selection which is more flexible and general in our algorithm with $M \in [0, 1]^d$.

The whole framework is jointly optimized, so that the mutual promotion between heterogeneity identification and invariant learning can be fully leveraged.

3.1. Implementation of $\mathcal{M}_p$

Here we introduce our invariant prediction module $\mathcal{M}_p$, which takes multiple environments training data $D = \{D^e\}_{e \in supp(\mathcal{E}_{tr})}$ as input, and outputs the corresponding invariant predictor f and the indices of invariant features M given current environments $\mathcal{E}_{tr}$. We combine feature selection with invariant learning under heterogeneous environments, which can select the features with stable/invariant correlations with the label across $\mathcal{E}_{tr}$. Specifically, the former module can select most informative features with respect to the loss function and latter module ensures the selected features are invariant. Their combination ensures $\mathcal{M}_p$ to select the most informative invariant features.

For invariant learning, we follow the variance penalty regularizer proposed in (Koyama & Yamaguchi, 2020) and simplify it in feature selection scenarios. The objective function of $\mathcal{M}_p$ with $M \in \{0, 1\}^d$ is:

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

$$\mathcal{L}_p(M \odot X, Y; \theta) = \mathbb{E}_{\mathcal{E}_{tr}}[\mathcal{L}^e] + \lambda \mathrm{trace}(\mathrm{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e)) \quad (5)$$

However, as the optimization of hard feature selection with binary mask M suffers from high variance, we use the soft feature selection with gates taking continuous value in $[0, 1]$. Specifically, following (Yamada et al., 2020), we approximate each element of $M = [m_1, \ldots, m_d]^T$ to clipped Gaussian random variable parameterized by $\mu = [\mu_1, \ldots, \mu_d]^T$ as

$$m_i = \max\{0, \min\{1, \mu_i + \epsilon\}\} \quad (6)$$

where ϵ is drawn from $\mathcal{N}(0, \sigma^2)$. With this approximation, the objective function with soft feature selection can be written as:

$$\mathcal{L}^e(\theta, \mu) = \mathbb{E}_{P^e}\mathbb{E}_M\left[\ell(M \odot X^e, Y^e; \theta) + \alpha \|M\|_0\right] \quad (7)$$

where M is a random vector with d independent variables m_i for $i \in [d]$. Under the approximation in Equation 6, $\|M\|_0$ is simply $\sum_{i \in [d]} P(m_i > 0)$ and can be calculated as $\|M\|_0 = \sum_{i \in [d]} \mathrm{CDF}(\mu_i/\sigma)$, where CDF is the standard Gaussian CDF. We formulate our objective as risk minimization problem:

$$\min_{\theta, \mu} \mathcal{L}_p(\theta; \mu) = \mathbb{E}_{\mathcal{E}_{tr}}[\mathcal{L}^e(\theta, \mu)] + \lambda \mathrm{trace}(\mathrm{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e)) \quad (8)$$

where

$$\mathcal{L}^e(\theta, \mu) = \mathbb{E}_{P^e}\mathbb{E}_M\left[\ell(M \odot X^e, Y^e; \theta) + \alpha \sum_{i \in [d]} \mathrm{CDF}(\mu_i/\sigma)\right] \quad (9)$$

Further, as for linear models, we simply approximate the regularizer $\mathrm{trace}(\mathrm{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e))$ by $\|\mathrm{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e) \odot M\|^2$. Then we obtain $\Phi(X)$ and $\Psi(X)$ when we obtain μ as well as M. Further in Section 4, we theoretically prove that the prediction module $\mathcal{M}_p$ is able to learn the MIP with respect to given environments $\mathcal{E}_{tr}$.

3.2. Implementation of $\mathcal{M}_c$

Notation. Ψ means the learned variant part $\Psi(X)$. Δ_K means K-dimension simplex. $f_\theta(\cdot)$ means the function f parameterized by θ.

The heterogeneity identification module $\mathcal{M}_c$ takes a single dataset as input, and outputs a multi-environment dataset partition for invariant prediction. We implement it with a clustering algorithm. As indicated in Theorem 2.3, the more diverse $P(Y|\Psi)$ for our generated environments, the better the invariance set $\mathcal{I}$ is. Therefore, we cluster the data points according to the relationship between Ψ and Y, for which we use $P(Y|\Psi)$ as the cluster centre. Note that Ψ is initialized as $\Psi(X) = X$ in our joint optimization.

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

$$h_j(\Psi, Y) = P_{\Theta_j}(Y|\Psi) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(Y - f_{\Theta_j}(\Psi))^2}{2\sigma^2}) \quad (10)$$

For the given $N = \sum_{e \in \mathcal{E}_{tr}} n_e$ empirical data samples $\mathcal{D} = \{\psi_i(x_i), y_i\}_{i=1}^N$, the empirical distribution is modeled as $\hat{P}_N = \frac{1}{N}\sum_{i=1}^N \delta_i(\Psi, Y)$ where

$$\delta_i(\Psi, Y) = \begin{cases} 1, & \text{if } \Psi = \psi_i \text{ and } Y = y_i \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

The target of our heterogeneous clustering 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. Therefore, the objective function of our heterogeneous clustering is:

$$\min_{Q \in \mathcal{Q}} D_{KL}(\hat{P}_N \| Q) \quad (12)$$

The above objective can be further simplified to:

$$\min_{\Theta, \mathbf{q}} \left\{ \mathcal{L}_c = -\frac{1}{N}\sum_{i=1}^N \log\left[\sum_{j=1}^K q_j h_j(\psi_i, y_i)\right]\right\} \quad (13)$$

As for optimization, we use EM algorithm to optimize the centre parameter Θ and the mixture weight $\mathbf{q}$. After optimizing equation 13, for building $\mathcal{E}_{tr}$, we assign each data point to environment $e_j \in \mathcal{E}_{tr}$ with probability:

$$P(e_j|\Psi, Y) = q_j h_j(\Psi, Y) / \left(\sum_{i=1}^K q_i h_i(\Psi, Y)\right) \quad (14)$$

In this way, $\mathcal{E}_{tr}$ is generated by $\mathcal{M}_c$.

4. Theoretical Analysis

In this section, we theoretically analyze our proposed Heterogeneous Risk Minimization (HRM) method. We first analyze our proposed invariant learning module $\mathcal{M}_p$, and then justify the existence of the positive feedback in our HRM.

Justification of $\mathcal{M}_p$ We prove that given training environments $\mathcal{E}_{tr}$, our invariant prediction model $\mathcal{M}_p$ can learn the maximal invariant predictor $\Phi(X)$ with respect to the corresponding invariance set $\mathcal{I}_{\mathcal{E}_{tr}}$.

Theorem 4.1. *Given $\mathcal{E}_{tr}$, the learned $\Phi(X) = M \odot X$ is the maximal invariant predictor of $\mathcal{I}_{\mathcal{E}_{tr}}$.*

Justification of the Positive Feedback The core of our HRM framework is the mechanism for $\mathcal{M}_c$ and $\mathcal{M}_p$ to mutual promote each other. Here we theoretically justify the existence of such positive feedback. In Assumption 2.1, we assume the invariance and sufficiency properties of the stable features Φ^* and assume the relationship between unstable part Ψ^* and Y can arbitrarily change. Here we make a more specific assumption on the heterogeneity across environments with respect to Φ^* and Ψ^*.

Assumption 4.1. *Assume the pooled training data is made up of heterogeneous data sources: $P_{tr} = \sum_{e\in\text{supp}(\mathcal{E}_{tr})} w_e P^e$. For any $e_i, e_j \in \mathcal{E}_{tr}, e_i \neq e_j$, we assume*

$$I^c_{i,j}(Y;\Phi^*|\Psi^*) \geq \max(I_i(Y;\Phi^*|\Psi^*), I_j(Y;\Phi^*|\Psi^*)) \quad (15)$$

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 $H^c_{i,j}[Y] = -\int p^{e_i}(y)\log p^{e_j}(y)dy$.

Here we would like to intuitively demonstrate this assumption. Firstly, 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. Then the cross mutual information $I_{i,j}(Y;\Phi^*)$ 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. Therefore, the R.H.S in equation 15 measures that, in environment e_i, how much prediction error can be reduced if we further add Φ^* for prediction rather than use only Ψ^*. And the L.H.S measures that, when using predictors trained in e_i to predict in e_j, how much prediction error can be reduced if we further add Φ^* for prediction rather than use only Ψ^*. Intuitively, Assumption 4.1 assumes that invariant feature Φ^* provides more information for predicting Y across environments than in one single environment, and correspondingly, the information provided by Ψ^* shrinks a lot across environments, which indicates that the relationship between variant feature Ψ^* and Y varies across environments. Based on this assumption, we first prove that the cluster centres are pulled apart as invariant feature is excluded from clustering.

Theorem 4.2. *For $e_i, e_j \in \text{supp}(\mathcal{E}_{tr})$, assume that $X = [\Phi^*, \Psi^*]^T$ satisfying Assumption 2.1, where Φ^* is invariant and Ψ^* variant. Then under Assumption 4.1, we have $\mathrm{D}_{\mathrm{KL}}(P^{e_i}(Y|X)\|P^{e_j}(Y|X)) \leq \mathrm{D}_{\mathrm{KL}}(P^{e_i}(Y|\Psi^*)\|P^{e_j}(Y|\Psi^*))$*

Theorem 4.2 indicates that the distance between cluster centres is larger when using variant features Ψ^* , therefore, it is more likely to obtain the desired heterogeneous environments, which explains why we use learned variant part $\Psi(X)$ for clustering. Finally, we provide the theorem for optimality guarantee for our HRM.

Theorem 4.3. *Under Assumption 2.1 and 4.1, for the proposed $\mathcal{M}_c$ and $\mathcal{M}_p$, we have the following conclusions: 1. Given environments $\mathcal{E}_{tr}$ such that $\mathcal{I}_{\mathcal{E}} = \mathcal{I}_{\mathcal{E}_{tr}}$, the learned $\Phi(X)$ by $\mathcal{M}_p$ is the maximal invariant predictor of $\mathcal{I}_{\mathcal{E}}$. 2. Given the maximal invariant predictor Φ^* of $\mathcal{I}_{\mathcal{E}}$, assume the pooled training data is made up of data from all environments in $\text{supp}(\mathcal{E})$, there exist one split that achieves the minimum of the objective function and meanwhile the invariance set regularized is equal to $\mathcal{I}_{\mathcal{E}}$.*

Intuitively, Theorem 4.3 proves that given one of the $\mathcal{M}_c$ and $\mathcal{M}_p$ optimal, the other is optimal, which validates the existence of the global optimal point of our algorithm.

5. Experiment

In this section, we validate the effectiveness of our method on simulation data and real-world data.

Baselines We compare our proposed HRM with the following methods:

- Empirical Risk Minimization(ERM): $\min_\theta \mathbb{E}_{P_0}[\ell(\theta;X,Y)]$
- Distributionally Robust Optimization(DRO (Sinha et al., 2018)): $\min_\theta \sup_{Q\in W(Q,P_0)\leq\rho} \mathbb{E}_Q[\ell(\theta;X,Y)]$
- Environment Inference for Invariant Learning(EIIL (Creager et al., 2020)):

$$\min_\Phi \max_u \sum_{e\in\mathcal{E}} \frac{1}{N_e}\sum_i u_i(e)\ell(w\odot\Phi(x_i),y_i) + \sum_{e\in\mathcal{E}} \lambda\|\nabla_{w|w=1.0}\frac{1}{N_e}\sum_i u_i(e)\ell(w\odot\Phi(x_i),y_i)\|_2 \quad (16)$$

- Invariant Risk Minimization(IRM (Arjovsky et al., 2019)) with environment $\mathcal{E}_{tr}$ labels:

$$\min_\Phi \sum_{e\in\mathcal{E}_{tr}} \mathcal{L}^e + \lambda\|\nabla_{w|w=1.0}\mathcal{L}^e(w\odot\Phi)\|^2 \quad (17)$$

Further, for ablation study, we also compare with HRM^s, which runs HRM for only one iteration without the feedback loop. Note that IRM is based on multiple training environments and we provide environment $\mathcal{E}_{tr}$ labels for IRM, while others do not need environment labels.

Evaluation Metrics To evaluate the prediction performance, we use Mean_Error defined as $\text{Mean\_Error} = \frac{1}{|\mathcal{E}_{test}|}\sum_{e\in\mathcal{E}_{test}}\mathcal{L}^e$, Std_Error defined as $\text{Std\_Error} = \sqrt{\frac{1}{|\mathcal{E}_{test}|-1}\sum_{e\in\mathcal{E}_{test}}(\mathcal{L}^e - \text{Mean\_Error})^2}$, which are mean and standard deviation error across $\mathcal{E}_{test}$ and $\text{Max\_Error} = \max_{e\in\mathcal{E}_{test}}\mathcal{L}^e$, which are mean error, standard deviation error and worst-case error across $\mathcal{E}_{test}$.

Imbalanced Mixture It is a natural phenomena that empirical data follow a power-law distribution, i.e. only a few environments/subgroups are common and the rest are rare (Shen et al., 2018; Sagawa et al., 2019; 2020). Therefore, we perform non-uniform sampling among different environments in training set.

5.1. Simulation Data

We design two mechanisms to simulate the varying correlations among covariates across environments, named by selection bias and anti-causal effect.

Selection Bias In this setting, the correlations between variant covariates and the target are perturbed through selection bias mechanism. According to Assumption 2.1, we assume $X = [\Phi^*, \Psi^*]^T \in \mathbb{R}^d$ and $Y = f(\Phi^*) + \epsilon$ and that $P(Y|\Phi^*)$ remains invariant across environments while $P(Y|\Psi^*)$ changes arbitrarily. For simplicity, we select data points according to a certain variable set $V_b \subset \Psi^*$:

$$\hat{P}(x) = \prod_{v_i \in V_b} |r|^{-5*|f(\phi^*) - sign(r)*v_i|} \tag{18}$$

where $|r| > 1$, $V_b \in \mathbb{R}^{n_b}$ and $\hat{P}(x)$ denotes the probability of point x to be selected. Intuitively, r eventually controls the strengths and direction of the spurious correlation between V_b and Y(i.e. if $r > 0$, a data point whose V_b is close to its y is more probably to be selected.). The larger value of $|r|$ means the stronger spurious correlation between V_b and Y, and $r \geq 0$ means positive correlation and vice versa. Therefore, here we use r to define different environments.

In training, we generate $sum = 2000$ data points, where $\kappa = 95\%$ points from environment e_1 with a predefined r and $1 - \kappa = 5\%$ points from e_2 with $r = -1.1$. In testing, we generate data points for 10 environments with $r \in [-3, -2.7, -2.3, \dots, 2.3, 2.7, 3.0]$. β is set to 1.0. We compare our HRM with ERM, DRO, EIIL and IRM for Linear Regression. We conduct extensive experiments with different settings on r, n_b and d. In each setting, we carry out the procedure 10 times and report the average results. The results are shown in Table 2.

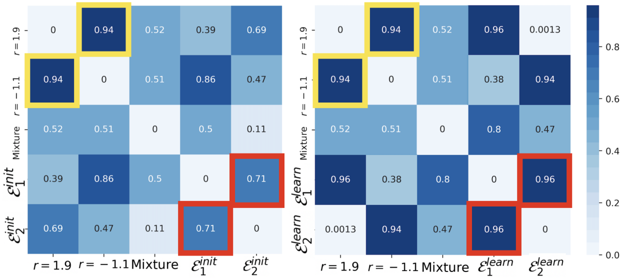


Figure 2. Visualization of differences between environments in scenario 1 in selection bias experiment($r = 1.9$). The left figure shows the initial clustering results using X, and the right one shows the learned $\mathcal{E}^{learn}$ using the learned variant part $\Psi(X)$.

From the results, we have the following observations and analysis: **ERM** suffers from the distributional shifts in testing and yields poor performance in most of the settings. **DRO** surprisingly has the worst performance, which we think is due to the over-pessimism problem (Frogner et al., 2019). **EIIL** has the similar performance with ERM, which indicates that its inferred environments cannot reveal the spurious correlations between Y and V_b. **IRM** performs much better than the above two baselines, however, as IRM depends on the available environment labels to work, it uses much more information than the other three methods. Compared to the three baselines, our **HRM** achieves nearly perfect performance with respect to average performance and stability, especially the variance of losses across environments is close to 0, which reflects the effectiveness of our heterogeneous clustering as well as the invariant learning algorithm. Furthermore, our HRM does not need environment labels, which verifies that our clustering algorithm can mine the latent heterogeneity inside the data and further shows our superiority to IRM.

Besides, we visualize the differences between environments using Task2Vec (Achille et al., 2019) in Figure 2, where larger value means the two environments are more heterogeneous. The pooled training data are mixture of environments with $r = 1.9$ and $r = -1.1$, the difference between whom is shown in yellow box. And the red boxes show differences between learned environments by HRM^s and HRM. The big promotion between $\mathcal{E}_{init}$ and $\mathcal{E}_{learn}$ verifies our HRM can exploit heterogeneity inside data as well as the existence of the positive feedback. Due to space limitation, results of varying sum, κ, n_b as well as experimental details are left to appendix.

Anti-causal Effect Inspired by (Arjovsky et al., 2019), we induce the spurious correlation by using anti-causal relationship from the target Y to the variant covariates Ψ^*. In this experiment, we assume $X = [\Phi^*, \Psi^*]^T \in \mathbb{R}^d$ and firstly sample Φ^* from mixture Gaussian distribution characterized as $\sum_{i=1}^k z_k \mathcal{N}(\mu_i, I)$ and the target $Y = \theta_\phi^T \Phi^* + \beta\Phi_1\Phi_2\Phi_3 + \mathcal{N}(0, 0.3)$. Then the spurious

Table 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**

Table 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**	**0.3200**	**0.317**	**0.326**	**0.330**	**0.333**	**0.335**

correlations between Ψ^* and Y are generated by anti-causal effect as

$$\Psi^* = \theta_\psi Y + \mathcal{N}(0, \sigma(\mu_i)^2) \tag{19}$$

where $\sigma(\mu_i)$ means the Gaussian noise added to Ψ^* depends on which component the invariant covariates Φ^* belong to. Intuitively, in different Gaussian components, the corresponding correlations between Ψ^* and Y are varying due to the different value of $\sigma(\mu_i)$. The larger the $\sigma(\mu_i)$ is, the weaker correlation between Ψ^* and Y. We use the mixture weight $Z = [z_1, \dots, z_k]^T$ to define different environments, where different mixture weights represent different overall strength of the effect Y on Ψ^*.

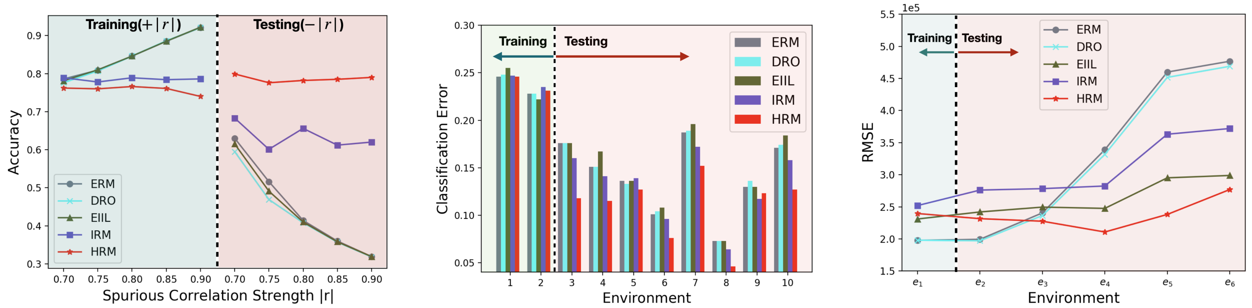
In this experiment, we set $\beta = 0.1$ and build 10 environments with varying σ and the dimension of Φ^*, Ψ^*, the first three for training and the last seven for testing. We run experiments for 10 times and the averaged results are shown in Table 3. **EIIL** achieves the best training performance with respect to prediction errors on training environments e_1, e_2, e_3, while its performances in testing are poor. **ERM** suffers from distributional shifts in testing. **DRO** seeks for over-considered robustness and performs much worse. **IRM** performs much better as it learns invariant representations with help of environment labels. **HRM** achieves nearly uniformly good performance in training environments as well as the testing ones, which validates the effectiveness of our method and proves its excellent generalization ability.

5.2. Real-world Data

We test our method on three real-world tasks, including car insurance prediction, people income prediction and house price prediction.

5.2.1. Settings

Car Insurance Prediction In this task, we use a real-world dataset for car insurance prediction (Kaggle). It is



(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.

Figure 3. Results of real-world datasets, including training and testing performance for five methods.

a classification task to predict whether a person will buy car insurance based on related information, such as vehicle damage, annual premium, vehicle age etc[1]. We impose selection bias mechanism on the correlation between the outcome (i.e. the label indicating whether buying insurance) and the sex attribute to simulate multiple environments. Specifically, we simulate different strengths $|r|$ of the spurious correlation between sex and target in training, and reverse the direction of such correlation in testing($+|r|$ in training and $-|r|$ in testing). For IRM, in each setting, we divide the training data into three training environments with $r_1 = 0.95, r_2 = 0.9, r_3 = -0.8$, and different overall correlation r corresponds to different numbers of data in e_1, e_2, e_3. We perform 5 experiments with varying r and the results in both training and testing are shown in Figure 3(a).

People Income Prediction In this task we use the Adult dataset (Dua & Graff, 2017) to predict personal income levels as above or below \$50,000 per year based on personal details. We split the dataset into 10 environments according to demographic attributes sex and race. In training phase, all methods are trained on pooled data including 693 points from environment 1 and 200 from environment 2, and validated on 100 sampled from both. For IRM, the ground-truth environment labels are provided. In testing phase, we test all methods on the 10 environments and report the mis-classification rate on all environments in Figure 3(b).

House Price Prediction In this experiment, we use a real-world regression dataset (Kaggle) of house sales prices from King County, USA[2]. The target variable is the transaction price of the house and each sample contains 17 predictive variables such as the built year of the house, number of bedrooms, and square footage of home, etc. We simulate different environments according to the built year of the house, since it is fairly reasonable to assume the correlations among covariates and the target may vary along time. Specifically, we split the dataset into 6 periods, where each period approximately covers a time span of two decades. All methods are trained on data from the first period($[1900, 1920)$) and test on the other periods. For IRM, we further divide the training data into two environments where $built\ year \in [1900, 1910)$ and $[1910, 1920)$ respectively. Results are shown in Figure 3(c).

5.2.2. Analysis

From the results of three real-world tasks, we have the following observations and analysis: **ERM** achieves high accuracy in training while performing much worse in testing, indicating its inability in dealing with OOD predictions. **DRO**'s performance is not satisfactory, sometimes even worse than ERM. One plausible reason is its over-pessimistic nature which leads to too conservative predictors. Comparatively, invariant learning methods perform better in testing. **IRM** performs better than ERM and DRO, which shows the usefulness of environment labels for OOD generalization and the possibility of learning invariant predictor from multiple environments. **EIIL** performs inconsistently across different tasks, possibly due to its instability of the environment inference method. In all tasks and almost all testing environments (16/18), **HRM** consistently achieves the best performances. HRM even outperforms IRM significantly in a unfair setting where we provide perfect environment labels for IRM. One one side, it shows the limitation of manually labeled environments. On the other side, it demonstrates that, relieving the dependence on environment labels, HRM can effectively uncover and fully leverage the intrinsic heterogeneity in training data for invariant learning.

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

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

6. Related Works

There are mainly two branches of methods for the OOD generalization problem, namely distributionally robust optimization (DRO) (Esfahani & Kuhn, 2018; Duchi & Namkoong, 2018; Sinha et al., 2018; Sagawa et al., 2019) and invariant learning (Arjovsky et al., 2019; Koyama & Yamaguchi, 2020; Chang et al., 2020; Creager et al., 2020). DRO methods propose to optimize the worst-case risk within an uncertainty set, which lies around the observed training distribution and characterizes the potential testing distributions. However, in real scenarios, to better capture the testing distribution, the uncertainty set should be pretty large, which also results in the over-pessimism problem of DRO methods(Hu et al., 2018; Frogner et al., 2019).

Realizing the difficulty of solving OOD generalization problem without any prior knowledge or structural assumptions, invariant learning methods assume the existence of causally invariant relationships between some predictors $\Phi(X)$ and the target Y. (Arjovsky et al., 2019) and (Koyama & Yamaguchi, 2020) propose to learning an invariant representation through multiple training environments. (Chang et al., 2020) also proposes to select features whose predictive relationship with the target stays invariant across environments. However, their effectiveness relies on the quality of the given multiple training environments, and the role of environments remains vague theoretically. Recently, (Creager et al., 2020) improves (Arjovsky et al., 2019) by relaxing its requirements for multiple environments. Specifically, (Creager et al., 2020) proposes a two-stage method, which firstly infers the environment division with a pre-provided biased model, and then performs invariant learning on the inferred environments. However, the two stages cannot be jointly optimized, and the environment division relies on the given biased model and lacks theoretical guarantees.

7. Discussions

In this work, we theoretically analyze the role of environments in invariant learning, and propose our HRM for joint heterogeneity identification and invariant prediction, which relaxes the requirements for environment labels and opens a new direction for invariant learning. To our knowledge, this is the first work to both theoretically and empirically analyze how the equality of multiple environments affects invariant learning. This paper mainly focuses on the raw variable level with the assumption of $X = [\Phi^*, \Psi^*]^T$, which is able to cover a broad spectrum of applications, e.g. healthcare, finance, marketing etc, where the raw variables are informative enough.

However, our work has some limitations, which we hope to improve in the future. Firstly, in order to achieve the mutual promotion, we should use the variant features Ψ^* for heterogeneity identification rather than the invariant ones. However, the process of invariant prediction continuously discards the variant features Ψ^* (for invariant features or representation), which makes it quite hard to recover the variant features. To overcome this, we focus on the simple setting where $X = [\Phi^*, \Psi^*]^T$, since we can directly obtain the variant features Ψ^* when having invariant features Φ^*. To further extend the power of HRM, we will consider to incorporate representation learning from X in future work. Secondly, our clustering algorithm in $\mathcal{M}_c$ lacks theoretical guarantees for its convergence. To the best of our knowledge, in order to theoretically analyze the convergence of a clustering algorithm, it is necessary to measure the distance between data points. However, our clustering algorithm takes models' parameters as cluster centers and aims to cluster data points (X, Y) according to their relationships between X and Y, whose dissimilarity cannot be easily measured, since the relationship is statistical magnitude and cannot be calculated individually. How to theoretically analyze the convergence property of such clustering algorithms remains unsolved.

8. Acknowledgements

This work was supported in part by National Key R&D Program of China (No. 2018AAA0102004, No. 2020AAA0106300), National Natural Science Foundation of China (No. U1936219, 61521002, 61772304), Beijing Academy of Artificial Intelligence (BAAI), and a grant from the Institute for Guo Qiang, Tsinghua University. Bo Li's research was supported by the Tsinghua University Initiative Scientific Research Grant, No. 2019THZWJC11; Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grant 2020AAA0108400 and 2020AAA01084020108403; Major Program of the National Social Science Foundation of China (21ZDA036).

References

Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C. C., Soatto, S., and Perona, P. Task2vec: Task embedding for meta-learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 6429–6438. IEEE, 2019. doi: 10.1109/ICCV.2019.00653. URL https://doi.org/10.1109/ICCV.2019.00653.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, pp. 0049124118782533, 2018.

Chang, S., Zhang, Y., Yu, M., and Jaakkola, T. S. Invariant rationalization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1448–1458. PMLR, 2020. URL http://proceedings.mlr.press/v119/chang20c.html.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *ICML Workshop on Uncertainty and Robustness*, 2020.

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

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

El Gamal, A. and Kim, Y.-H. Network information theory. *Network Information Theory*, 12 2011. doi: 10.1017/CBO9781139030687.

Esfahani, P. M. and Kuhn, D. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Math. Program.*, 171(1-2):115–166, 2018. doi: 10.1007/s10107-017-1172-1. URL https://doi.org/10.1007/s10107-017-1172-1.

Frogner, C., Claici, S., Chien, E., and Solomon, J. Incorporating unlabeled data into distributionally robust learning. *arXiv preprint arXiv:1912.07729*, 2019.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., and Schölkopf, B. Domain adaptation with conditional transferable components. In Balcan, M. and Weinberger, K. Q. (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 2839–2848. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/gong16.html.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2034–2042. PMLR, 2018. URL http://proceedings.mlr.press/v80/hu18a.html.

Koyama, M. and Yamaguchi, S. Out-of-distribution generalization with maximal invariant predictor. *CoRR*, abs/2008.01883, 2020. URL https://arxiv.org/abs/2008.01883.

Kuang, K., Xiong, R., Cui, P., Athey, S., and Li, B. Stable prediction with model misspecification and agnostic distribution shift. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 4485–4492. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/5876.

Kukar, M. Transductive reliability estimation for medical diagnosis. *Artificial Intelligence in Medicine*, 29(1-2): 81–106, 2003.

Rojas-Carulla, M., Schlkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. *Stats*, 2015.

Rudin, C. and Ustun, B. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, 48(5):449–466, 2018.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. 2020.

Shen, Z., Cui, P., Kuang, K., Li, B., and Chen, P. Causally regularized learning with agnostic data selection bias. In *2018 ACM Multimedia Conference*, 2018.

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. *International Conference on Learning Representations*, 2018.

Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. Feature selection using stochastic gates. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10648–10659. PMLR, 2020. URL http://proceedings.mlr.press/v119/yamada20a.html.