Appendix

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

A. Additional Simulation Results and Details

Selection Bias In this setting, the correlations among covariates are perturbed through selection bias mechanism. According to assumption 2.1, we assume $X = [\Phi^*, \Psi^*]^T \in \mathbb{R}^d$ and $\Phi^* = [\Phi_1^*, \Phi_2^*, \dots, \Phi_{n_\phi}^*]^T \in \mathbb{R}^{n_\phi}$ is independent from $\Psi^* = [\Psi_1^*, \Psi_2^*, \dots, \Psi_{n_\psi}^*] \in \mathbb{R}^{n_\psi}$ while the covariates in Φ^* are dependent with each other. We assume $Y = f(\Phi^*) + \epsilon$ and $P(Y|\Phi^*)$ remains invariant across environments while $P(Y|\Psi^*)$ can arbitrarily change.

Therefore, we generate training data points with the help of auxiliary variables $Z \in \mathbb{R}^{n_\phi+1}$ as following:

$$Z_1, \dots, Z_{n_\phi+1} \overset{iid}{\sim} \mathcal{N}(0, 1.0) \tag{1}$$

$$\Psi_1^*, \dots, \Psi_{n_\psi}^* \overset{iid}{\sim} \mathcal{N}(0, 1.0) \tag{2}$$

$$\Phi_i^* = 0.8 * Z_i + 0.2 * Z_{i+1} \quad for \;\; i = 1, \dots, n_\phi \tag{3}$$

To induce model misspecification, we generate Y as:

$$Y = f(\Phi^*) + \epsilon = \theta_\phi * (\Phi^*)^T + \beta * \Phi_1^*\Phi_2^*\Phi_3^* + \epsilon \tag{4}$$

where $\theta_\phi = [\frac{1}{2}, -1, 1, -\frac{1}{2}, 1, -1, \dots] \in \mathbb{R}^{n_\phi}$, and $\epsilon \sim \mathcal{N}(0, 0.3)$. As we assume that $P(Y|\Phi^*)$ remains unchanged while $P(Y|\Psi^*)$ can vary across environments, we design a data selection mechanism to induce this kind of distribution shifts. For simplicity, we select data points according to a certain variable set $V_b \subset \Psi^*$:

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

$$\mu \sim Uni(0, 1) \tag{6}$$

$$M(r; (x, y)) = \begin{cases} 1, & \mu \le \hat{P} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $|r| > 1$ and $V_b \in \mathbb{R}^{n_b}$. Given a certain r, a data point (x, y) is selected if and only if $M(r; (x, y)) = 1$ (i.e. if $r > 0$, a data point whose V_b is close to its Y is more probably 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 $r \ge 0$ means positive correlation and vice versa. Therefore, here we use r to define different environments.

In training, we generate sum data points, where $\kappa \cdot sum$ points from environment e_1 with a predefined r and $(1 - \kappa)sum$ points from e_2 with $r = -1.1$. In testing, we generate data points for 10 environments with $r \in [-3, -2, -1.7, \dots, 1.7, 2, 3]$. β is set to 1.0.

Apart from the two scenarios in main body, we also conduct scenario 3 and 4 with varying κ, n and n_b respectively.

Anti-Causal Effect Inspired by (Arjovsky et al., 2019), in this setting, we introduce the spurious correlation by using anti-causal relationship from the target Y to the variant covariates Ψ^*.

We assume $X = [\Phi^*, \Psi^*]^T \in \mathbb{R}^d$ and $\Phi^* = [\Phi_1^*, \Phi_2^*, \dots, \Phi_{n_\phi}^*]^T \in \mathbb{R}^{n_\phi}$, $\Psi^* = [\Psi_1^*, \Psi_2^*, \dots, \Psi_{n_\psi}^*] \in \mathbb{R}^{n_\psi}$ Data Generation process is as following:

$$\Phi^* \sim \sum_{i=1}^{k} z_k \mathcal{N}(\mu_i, I) \tag{8}$$

$$Y = \theta_\phi^T \Phi^* + \beta \Phi_1^*\Phi_2^*\Phi_3^* + \mathcal{N}(0, 0.3) \tag{9}$$

$$\Psi^* = \theta_\psi Y + \mathcal{N}(0, \sigma(\mu_i)^2) \tag{10}$$

where $\sum_{i=1}^k z_i = 1$ & $z_i >= 0$ is the mixture weight of k Gaussian components, $\sigma(\mu_i)$ means the Gaussian noise added to Ψ^* depends on which component the invariant covariates Φ^* belong to and $\theta_\psi \in \mathbb{R}^{n_\psi}$. 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

[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.

Table 1. Results in selection bias simulation experiments of different methods with varying sample size sum, ratio κ and variant dimensions n_b of training data, and each result is averaged over ten times runs.

Scenario 3: varying ratio κ and sample size sum $(d = 10, r = 1.9, n_b = 1)$									
κ, n	$\kappa = 0.90, sum = 1000$			$\kappa = 0.95, sum = 2000$			$\kappa = 0.975, sum = 4000$		
Methods	Mean_Error	Std_Error	Max_Error	Mean_Error	Std_Error	Max_Error	Mean_Error	Std_Error	Max_Error
ERM	0.477	0.061	0.530	0.510	0.108	0.608	0.547	0.150	0.687
DRO	0.480	0.107	0.597	0.512	0.111	0.625	0.608	0.227	0.838
EIIL	0.476	0.063	0.529	0.507	0.102	0.613	0.539	0.148	0.689
IRM(with $\mathcal{E}_{tr}$ label)	0.455	0.015	0.471	0.456	0.015	0.472	0.456	0.015	0.472
HRM	**0.450**	**0.010**	**0.461**	**0.447**	**0.011**	**0.465**	**0.447**	**0.010**	**0.463**
Scenario 4: varying variant dimension n_b $(d = 10, sum = 2000, \kappa = 0.95, r = 1.9, n_b = 1)$									
n_b	$n_b = 1$			$n_b = 3$			$n_b = 5$		
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	0.468	0.110	0.583	0.445	0.112	0.567
DRO	0.512	0.111	0.625	0.515	0.107	0.617	0.454	0.122	0.577
EIIL	0.520	0.111	0.613	0.469	0.111	0.581	0.454	0.100	0.557
IRM(with $\mathcal{E}_{tr}$ label)	0.456	0.015	0.472	0.432	0.014	0.446	0.414	0.061	0.475
HRM	**0.447**	**0.011**	**0.465**	**0.413**	**0.012**	**0.431**	**0.402**	**0.057**	**0.462**

build 10 environments with varying σ and the dimension of Φ^*, Ψ^*, the first three for training and the last seven for testing. Specifically, we set $\beta = 0.1, \mu_1 = [0,0,0,1,1]^T, \mu_2 = [0,0,0,1,-1]^T, \mu_2 = [0,0,0,-1,1]^T, \mu_4 = \mu_5 = \cdots = \mu_{10} = [0,0,0,-1,-1]^T$, $\sigma(\mu_1) = 0.2, \sigma(\mu_2) = 0.5, \sigma(\mu_3) = 1.0$ and $[\sigma(\mu_4), \sigma(\mu_5), \ldots, \sigma(\mu_{10})] = [3.0, 5.0, \ldots, 15.0]$. θ_ϕ, θ_ψ are randomly sampled from $\mathcal{N}(1, I)$ and $\mathcal{N}(0.5, 0.1I)$ respectively. We run experiments for 10 times and average the results.

B. Proofs

B.1. Proof of Theorem 2.1

First, we would like to prove that a random variable satisfying assumption 2.1 is MIP.

Theorem B.1. *A representation $\Phi^* \in \mathcal{I}$ satisfying assumption 2.1 is the maximal invariant predictor.*

Proof. $\rightarrow$: To prove $\Phi^* = \arg\min_{Z\in\mathcal{I}} I(Y;Z)$. If Φ^* is not the maximal invariant predictor, assume $\Phi' = \arg\max_{Z\in\mathcal{I}} I(Y;Z)$. Using functional representation lemma, consider (Φ^*, Φ'), there exists random variable Φ_{extra} such that $\Phi^{'} = \sigma(\Phi^*, \Phi_{extra})$ and $\Phi^* \perp \Phi_{extra}$. Then $I(Y;\Phi^{'}) = I(Y;\Phi^*,\Phi_{extra}) = I(f(\Phi^*);\Phi^*,\Phi_{extra}) = I(f(\Phi^*);\Phi^*)$.

$\leftarrow$: To prove the maximal invariant predictor Φ^* satisfies the sufficiency property in assumption 2.1.

The converse-negative proposition is :

$$Y \neq f(\Phi^*) + \epsilon \rightarrow \Phi^* \neq \arg\max_{Z\in\mathcal{I}} I(Y;Z) \tag{11}$$

Suppose $Y \neq f(\Phi^*) + \epsilon$ and $\Phi^* = \arg\max_{Z\in\mathcal{I}} I(Y;Z)$, and suppose $Y = f(\Phi^{'}) + \epsilon$ where $\Phi^{'} \neq \Phi^*$. Then we have:

$$I(f(\Phi^{'});\Phi^*) \leq I(f(\Phi^{'});\Phi^{'}) \tag{12}$$

Therefore, $\Phi^{'} = \arg\max_{Z\in\mathcal{I}} I(Y;Z)$ $\square$

Then we provide the proof of theorem 2.1.

Theorem B.2. *Let g be a strictly convex, differentiable function and let D be the corresponding Bregman Loss function. Let Φ^* is the maximal invariant predictor with respect to $I_\mathcal{E}$, and put $h^*(X) = \mathbb{E}_Y[Y|\Phi^*]$. Under assumption 2.2, we have:*

$$h^* = \arg\min_h \sup_{e\in\text{supp}(\mathcal{E})} \mathbb{E}[D(h(X),Y)|e] \tag{13}$$

Proof. Firstly, according to theorem B.1, Φ^* satisfies assumption 2.1. Consider any function h, we would like to prove that for each distribution $P^e(e \in \mathcal{E})$, there exists an environment e' such that:

$$\mathbb{E}[D(h(X),Y)|e'] \geq \mathbb{E}[D(h^*(X),Y)|e] \tag{14}$$

For each $e \in \mathcal{E}$ with density $([\Phi,\Psi],Y) \mapsto P(\Phi,\Psi,Y)$, we construct environment e' with density $Q(\Phi,\Psi,Y)$ that satisfies: (omit the superscript $*$ of Φ and Ψ for simplicity)

$$Q(\Phi,\Psi,Y) = P(\Phi,Y)Q(\Psi) \tag{15}$$

Note that such environment e' exists because of the heterogeneity property assumed in assumption 2.2. Then we have:

$$\int D(h(\phi,\psi),y)q(\phi,\psi,y)d\phi d\psi dy \tag{16}$$
$$= \int_\psi \int_{\phi,y} D(h(\phi,\psi),y)p(\phi,y)q(\psi)d\phi dy d\psi \tag{17}$$
$$= \int_\psi \int_{\phi,y} D(h(\phi,\psi),y)p(\phi,y)d\phi dy q(\psi)d\psi \tag{18}$$
$$\geq \int_\psi \int_{\phi,y} D(h^*(\phi,\psi),y)p(\phi,y)d\phi dy q(\psi)d\psi \tag{19}$$
$$= \int_\psi \int_{\phi,y} D(h^*(\phi),y)p(\phi,y)d\phi dy q(\psi)d\psi \tag{20}$$
$$= \int_{\phi,y} D(h^*(\phi),yp(\phi,y)d\phi dy \tag{21}$$
$$= \int_{\phi,\psi,y} D(h^*(\phi),y)p(\phi,\psi,y)d\phi d\psi dy \tag{22}$$
$$\tag{23}$$

□

B.2. Proof of Theorem 2.2

Theorem B.3. $\mathcal{I}_\mathcal{E} \subseteq \mathcal{I}_{\mathcal{E}_{tr}}$

Proof. Since $\mathcal{E}_{tr} \subseteq \mathcal{E}$, then for any $S \in \mathcal{I}_\mathcal{E}$, $S \in \mathcal{I}_{\mathcal{E}_{tr}}$. □

B.3. Proof of Theorem 2.3

Theorem B.4. *Given set of environments* $\mathrm{supp}(\hat{\mathcal{E}})$*, 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* $e \in \mathrm{supp}(\hat{\mathcal{E}})$*, the invariance set constrained by* $\mathrm{supp}(\hat{\mathcal{E}}) \cup \{e_{new}\}$ *is equal to* $\mathcal{I}_{\hat{\mathcal{E}}}$.

Proof. Denote the invariance set with respect to $\mathrm{supp}(\hat{\mathcal{E}} \cup \{e_{new}\})$ as $\mathcal{I}_{new}$, it is easy to prove that $\forall S \in \mathcal{I}_{\hat{\mathcal{E}}}$, we have $S \in \mathcal{I}_{new}$, since the newly-added environment cannot exclude any variables from the original invariance set. □

B.4. Proof of Theorem 4.1

Theorem B.5. *Given* $\mathcal{E}_{tr}$*, the learned* $\Phi(X) = M \odot X$ *is the maximal invariant predictor of* $\mathcal{I}_{\mathcal{E}_{tr}}$.

Proof. The objective function for $\mathcal{M}_p$ is

$$\mathcal{L}_p(M \odot X, Y; \theta) = \mathbb{E}_{\mathcal{E}_{tr}}[\mathcal{L}^e] + \lambda \mathrm{trace}(\mathrm{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e)) \tag{24}$$

Here we prove that the minimum of objective function can be achieved when $\Phi(X) = M \odot X$ is the maximal invariant predictor. According to theorem B.1, $\Phi(X)$ satisfies assumption 2.1, which indicates that $P^e(Y|\Phi(X))$ stays invariant.

From the proof in C.2 in (Koyama & Yamaguchi, 2020), $I(Y;\mathcal{E}|\Phi(X)) = 0$ indicates that $\mathrm{trace}(\mathrm{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e)) = 0$.

Further, from the sufficiency property, the minimum of $\mathcal{L}^e$ is achieved with $\Phi(X)$. Therefore, $\mathbb{E}_{\mathcal{E}_{tr}}[\mathcal{L}^e] + \lambda\mathrm{trace}(\mathrm{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e))$ reaches the minimum with $\Phi(X)$ being the MIP.($\lambda \geq 0$) □

B.5. Proof of Theorem 4.2

Theorem B.6. *For* $e_i, e_j \in \mathrm{supp}(\mathcal{E}_{tr})$*, assume that* $X = [\Phi^*, \Psi^*]^T$ *satisfying Assumption 2.1, where* Φ^* *is invariant and* Ψ^* *variant. Then under Assumption 4.1, we have* $\mathrm{D_{KL}}(P^{e_i}(Y|X)\|P^{e_j}(Y|X)) \leq \mathrm{D_{KL}}(P^{e_i}(Y|\Psi^*)\|P^{e_j}(Y|\Psi^*))$

Proof.

$$\mathrm{D_{KL}}(P^{e_i}(Y|X)\|P^{e_j}(Y|X)) \tag{25}$$
$$= \mathrm{D_{KL}}(P^{e_i}(Y|\Phi^*,\Psi^*)\|P^{e_j}(Y|\Phi^*,\Psi^*)) \tag{26}$$
$$= \int\int\int p_i(y,\phi,\psi)\log\left[\frac{p_i(y|\phi,\psi)}{p_j(y|\phi,\psi)}\right] dy d\phi d\psi \tag{27}$$

Therefore, we have

$$\mathrm{D_{KL}}(P^{e_i}(Y|\Psi)\|P^{e_j}(Y|\Psi)) - \mathrm{D_{KL}}(P^{e_i}(Y|X)\|P^{e_j}(Y|X)) \tag{28}$$
$$= \int\int\int p_i(y,\phi,\psi)\left(\log\frac{p_i(y|\psi)}{p_j(y|\psi)} - \log\frac{p_i(y|\phi,\psi)}{p_j(y|\phi,\psi)}\right) dy d\phi d\psi \tag{29}$$
$$= \int\int\int p_i(y,\phi,\psi)\left(\log\frac{p_i(y|\psi)}{p_i(y|\phi,\psi)} - \log\frac{p_j(y|\psi)}{p_j(y|\phi,\psi)}\right) dy d\phi d\psi \tag{30}$$
$$= I^c_{i,j}(Y;\Phi^*|\Psi^*) - I_i(Y;\Phi^*|\Psi^*) \tag{31}$$

Therefore, we have

$$\mathrm{D_{KL}}(P^{e_i}(Y|X)\|P^{e_j}(Y|X)) \leq \mathrm{D_{KL}}(P^{e_i}(Y|\Psi^*)\|P^{e_j}(Y|\Psi^*)) \tag{32}$$

□

B.6. Proof of Theorem 4.3

Theorem B.7. *Under Assumption 2.1 and 2.2, 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* $\mathrm{supp}(\mathcal{E})$*, then the invariance set* $\mathcal{I}_{\mathcal{E}_{tr}}$ *regularized by learned environments* $\mathcal{E}_{tr}$ *is equal to* $\mathcal{I}_\mathcal{E}$.

Proof. For 1, according to theorem B.5, the learned $\Phi(X)$ by $\mathcal{M}_p$ is the maximal invariant predictor of $\mathcal{I}_{\mathcal{E}_{tr}}$. Therefore, if $\mathcal{I}_\mathcal{E} = \mathcal{I}_{\mathcal{E}_{tr}}$, $\Phi(X)$ is the real maximal invariant predictor.

For 2, assume that $P_{train}(X,Y) = \sum_{e\in\mathcal{E}} w_e P^e(X,Y)$, we would like to prove that $\mathrm{D}_{KL}(P_{train}(Y|\Psi^*)\|Q)$ reaches minimum when the components in the mixture distribution Q corresponds to distributions for $e \in \mathcal{E}$. Since the learned $\Phi(X)$ by $\mathcal{M}_p$ is the maximal invariant predictor of $\mathcal{I}_\mathcal{E}$, the corresponding $\Psi(X)$ is exactly the $\Psi^*(X)$. Then taking $Q^* = \sum_{e\in\mathcal{E}} w_e P^e(Y|\Psi^*)$, we have $\forall Q \in \mathcal{Q}$,

$$\mathrm{D}_{\mathrm{KL}}(P_{train}(Y|\Psi^*)\|Q^*) \le \mathrm{D}_{\mathrm{KL}}(P_{train}(Y|\Psi^*)\|Q) \tag{33}$$

Therefore, the components in Q^* correspond to P^e for $e \in \mathcal{E}$, which makes $\mathcal{I}_{\mathcal{E}_{tr}}$ approaches to $\mathcal{I}_\mathcal{E}$. $\square$

References

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

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.