Covariate-Shift Generalization via Random Sample Weighting

Yue He[1], Xinwei Shen[2], Renzhe Xu[1], Tong Zhang[3], Yong Jiang[1], Wenchao Zou[4], Peng Cui[1]*

[1]Tsinghua University
[2]ETH Zürich
[3]The Hong Kong University of Science and Technology
[4]Siemens
heyue18@mails.tsinghua.edu.cn, xinwei.shen@stat.math.ethz.ch, xrz199721@gmail.com, tongzhang@tongzhang-ml.org, jiangyong@sz.tsinghua.edu.cn, wenchao.zou@siemens.com, cuip@tsinghua.edu.cn

Abstract

Shifts in the marginal distribution of covariates from training to the test phase, named covariate-shifts, often lead to unstable prediction performance across agnostic testing data, especially under model misspecification. Recent literature on invariant learning attempts to learn an invariant predictor from heterogeneous environments. However, the performance of the learned predictor depends heavily on the availability and quality of provided environments. In this paper, we propose a simple and effective non-parametric method for generating heterogeneous environments via Random Sample Weighting (RSW). Given the training dataset from a single source environment, we randomly generate a set of covariate-determining sample weights and use each weighted training distribution to simulate an environment. We theoretically show that under appropriate conditions, such random sample weighting can produce sufficient heterogeneity to be exploited by common invariance constraints to find the invariant variables for stable prediction under covariate shifts. Extensive experiments on both simulated and real-world datasets clearly validate the effectiveness of our method.

1 Introduction

Despite the great success of machine learning, it often assumes the independent and identically distributed (IID) training and test data, encouraging the models to minimize the empirical training error. However, the IID assumption is fragile in practice since distribution shifts often occur due to various reasons such as sample selection bias. Such distribution shifts lead to poor performance of traditional algorithms, especially when the test distribution is unknown. The risk is more critical in high-stake scenarios such as financial analysis (Wong, Hryniowski, and Wang 2020; Xu et al. 2022a), medical diagnosis (Kukar 2003), and criminal justice (Rudin and Ustun 2018). To guarantee the good generalization ability of a model on data drawn out-of-distribution (OOD) (Koh et al. 2021), the OOD generalization problem has been intensively studied in recent years.

Among types of distribution shifts, covariate-shift (Ben-David et al. 2010) is the most common one, where the marginal distribution of covariates $P(x)$ changes from the training to the test phase, while the conditional distribution of the response variable $P(y|x)$ remains unchanged. If a model is correctly specified and $P(y|x)$ can be learned precisely, its performance will not be affected by covariate-shifts. However, the model misspecification problem is usually inevitable in real applications. Therefore, here we focus on the covariate-shift generalization problem, i.e., how to improve the generalization ability of a model under model misspecification and unknown covariate-shifts.

There are several strands of literature related to the target problem. Given multiple heterogeneous environments, invariant learning methods and domain generalization methods are proposed to learn a domain-agnostic model (Ghifary et al. 2015; Li et al. 2017; Ganin et al. 2017) or invariant representation (Muandet, Balduzzi, and Schölkopf 2013; Lee, Kim, and Lee 2021; Krueger et al. 2021). However, their performance depends heavily on the availability and quality of heterogeneous environments (Ahuja et al. 2021), which are difficult to guarantee in real applications. In contrast, stable learning algorithms (Shen et al. 2020; Kuang et al. 2020; Xu et al. 2022b; Zhang et al. 2021) require only one environment. Enlightened by the ideas of causal inference, they tend to identify the causal variables by learning sample weights to make all covariates mutually independent and exploit the invariance property of causal variables for covariate-shift generalization. But the underlying assumptions of causal inference mean that their performance is largely affected by the sample size and the inner heterogeneity of the training data (Rosenbaum and Rubin 1983). Some recent studies stand in the middle. HRM (Liu et al. 2021b) and EIIL (Creager, Jacobsen, and Zemel 2021) put forward to generate heterogeneous environments adaptively from a mixture of data supposed to come from multiple environments without environment labels, their environment partition may bring negative effects unconsciously if the multi-environment hypothesis is violated.

In this paper, we combine the merits of stable learning and invariant learning. We suppose the observed variables X are composed of an invariant component S and a variant component V, i.e., $P(y|s)$ keeps unchanged while $P(y|v)$ changes under covariate-shifts. We found that by assigning a set of random weights on samples as long as the weights do not change $P(y|x)$, the $P(y|v)$ will be changed while $P(y|s)$ keeps unchanged. This means that such random weighting

*Corresponding Author

is an adequate way to simulate environments for differentiating S and V. We further extend this point by theoretical analysis that such random weighting can provide sufficient heterogeneity to be utilized by some common invariance constraints in finding the invariant variables, such as variance penalization. Motivated by the theory, we propose a novel Random Sample Weighting (RSW) method to generate covariate-determining random weights, and integrate it into each optimization step of invariant learning methods.

The main contributions of this paper are as follows:

- We theoretically analyze the risk of covariate-shift generalization under model misspecification and the role of random weighting, particularly how it affects estimation.
- Motivated by the theoretical findings, we further propose a simple non-parametric method that simulates a large number of heterogeneous environments via random covariate-determining sample weights for learning stable prediction models with auxiliary invariance constraints.
- We carry out extensive experiments to clearly validate the effectiveness of our method.

2 Preliminary

2.1 Problem Setup

Assumption 2.1 (Covariate Shift)**.** The test distribution $P^{te}(x, y)$ differs from the training distribution $P^{tr}(x, y)$ in the shift of covariate distribution only, i.e.,

$$P^{te}(x, y) = P^{te}(x)P^{tr}(y|x), \tag{1}$$

where $P^{te}(x)$ has the same support with $P^{tr}(x)$.

Problem 2.1 (Covariate Shift Generalization)**.** Given the samples $\{(x_i, y_i)\}_{i=1}^N$ drawn from the training distribution $P^{tr}(x, y)$, the goal of covariate shift generalization problem is to learn a prediction model that performs stably in an agnostic test distribution $P^{te}(x)$ that satisfies Assumption 2.1.

Because the observed variables $X \in \mathbb{R}^D$ usually contain the invariant and variant components, here we define the invariant variable set S in Definition 2.1.

Definition 2.1 (Invariant Variable Set)**.** The subset $S \subset X$ is an invariant variable set iff S satisfies Equation (2) and none of its proper subsets satisfy Equation (2), where $V = X \setminus S$.

$$Y \perp V \mid S. \tag{2}$$

In contrast, we call V the variant variable set. According to Equation (1) and Equation (2), it is easy to state: $P^{tr}(y|s) = P^{tr}(y|x)$ is invariant across all the potential distributions.

Notation Throughout the paper, we use upper-cased letters X, Y to denote random variables or vectors, lower-cased letters x, y to denote their realizations/observations, and bold capital letters $\mathbf{X}, \mathbf{Y}$ to denote the matrices or vectors containing all the observations of X, Y. All distributions are assumed to be absolutely continuous with respect to the Lebesgue measure unless stated otherwise.

2.2 Empirical Risk Minimization

Empirical Risk Minimization (ERM) minimizes the average empirical errors of training samples $\{x_i, y_i\}_{i=1}^N$. Given a loss function in form of $\mathcal{L}(x, y; \theta)$ (e.g. $(\theta^\top x - y)^2$ used in least squares regression) where θ denotes the parameter set, ERM minimizes the following objective:

$$\min_\theta \ \mathbb{E}_{P^{tr}}[\mathcal{L}(X, Y; \theta)] \approx \frac{1}{N}\sum_{i=1}^N \mathcal{L}(x_i, y_i; \theta). \tag{3}$$

2.3 Importance Sampling Weights

Importance Sampling Weight (Hassanpour and Greiner 2019) is defined as $w(x, y) = P^{te}(x, y)/P^{tr}(x, y)$. Equation (4) illustrates that it can approximate the expected test error using the weighted empirical errors of training data.

$$\begin{aligned}\mathbb{E}_{P^{te}}[\mathcal{L}(X, Y; \theta)] &= \int \frac{P^{te}(x, y)}{P^{tr}(x, y)}\mathcal{L}(x, y; \theta)P^{tr}(x, y)dxdy \\ &\approx \frac{1}{N}\sum_{i=1}^N w(x_i, y_i)\mathcal{L}(x_i, y_i; \theta).\end{aligned} \tag{4}$$

2.4 Invariant Learning

Invariant Learning aims to capture an representation $\Phi(X)$, so that $P(Y|\Phi(X))$ keeps invariant across all potential distributions. Notably, Invariant Risk Minimization (IRM) (Arjovsky et al. 2019) proposes to minimize the following objective, given data collected from multiple training environments $\mathcal{E}_{tr}$.

$$\min \sum_{e\in\mathcal{E}_{tr}} \mathcal{L}^e(\Phi(X^e), Y^e; \theta) + \lambda \cdot \mathbb{D}(\theta, \Phi, e), \tag{5}$$

where $\mathbb{D}(\theta, \Phi, e) = \|\theta - \theta^e_\Phi\|^2$ is an invariance constraint describing the distance of learnt parameter Φ and the best solution Φ^e in e-th environment. Note that

$$\sum_{e\in\mathcal{E}_{tr}} \mathbb{D}(\theta, \Phi, e) \geq \left\| \frac{\sum_{e\in\mathcal{E}_{tr}} \theta^e_\Phi}{|\mathcal{E}_{tr}|} - \theta^e_\Phi \right\|^2 = \mathrm{Var}(\theta^e_\Phi) \tag{6}$$

which suggests that the optimal $\Phi(X)$ leads to the minimal variance of parameter θ^e_Φ across environments E_{tr}. This enlightens us a discriminative property of $\Phi(X)$.

3 Algorithm

3.1 Theoretical Motivation

In this section, we present the theoretical results that motivate the algorithm for covariate-shift generation via random sample weighting proposed in Section 3.2. We consider the problem of linear regression with model misspecification. Specifically, denote the covariates by $X = (S, V)$ and the response variable by Y, where S is an invariant variable set according to Definition 2.1. Without loss of generality, we assume both X and Y are standardized with zero means and

Algorithm 1: Covariate-shift Generalization via Random Sample Weighting (RSW)

Input: the training dataset $\{(x_i, y_i)\}_{i=1}^N$, the number of sample weighting E, and a predefined weighting function type $w(x; B)$ and distribution P_B of parameters B.
Initialize the parameter set θ of stable prediction model.
repeat
  Sample $\{B_1, \cdots, B_E\} \sim P_B$.
  Calculate $\mathcal{L} = 1/E \sum_{e=1}^{E} \sum_{i=1}^{N} w(x_i; B_e) \cdot \mathcal{L}(M \odot x_i, y_i; \theta) + \lambda \cdot \| \bigtriangledown_\theta \mathcal{L}^e(\theta) \odot M\|^2 + \alpha \cdot \|M\|_0$.
  Optimize $\theta \leftarrow \theta - \eta \cdot \bigtriangledown \mathcal{L}$.
until convergence
return: the stable prediction model with parameters θ.

unit variances. For simplicity, throughout this section, we consider the scalar case with $S, V \in \mathbb{R}$ being random variables, although most discussion can be readily extended to the vector case. The true model is given by

$$Y = \beta_S S + \beta_V V + g(S) + \epsilon,$$

where $\beta_V = 0$ according to Equation (2), $g(\cdot)$ is a nonlinear function representing the model misspecification, and $\epsilon \sim \mathcal{N}(0, 1)$ is the random error. Let $P(x, y)$ denote the distribution induced by the true model.

We use a linear model class which is misspecified due to the existence of $g(S)$. Let $\beta^* = \arg\min_b \mathbb{E}_P[Y - b^\top X]^2$ be the population ordinary least squares (OLS) solution. Turning to the sample case, we consider N IID training samples $\{x_i, y_i\}_{i=1}^N$ from P. Then the empirical OLS (ERM) solution is given by $\hat{\beta} = \arg\min_b \frac{1}{N} \sum_{i=1}^{N} (y_i - b^\top x_i)^2$.

We first identify the problem that under model misspecification, the ERM solution tends to include the variant components V, leading to a variant model vulnerable to distribution shifts. The following proposition illustrates the problem under a simple case where covariates S and V have zero correlation coefficient (for simplicity) but are correlated through the misspecified term $g(S)$ (corresponding to the spurious correlation). Since $\hat{\beta}$ is consistent (i.e., converging in probability) to β^*, Proposition 3.1 implies that $\hat{\beta}_V \neq 0$ asymptotically with a high probability. The proof is given in Appendix.

Proposition 3.1. *Assume* $\mathbb{E}[SV] = 0$ *and* $\mathbb{E}[g(S)V] \neq 0$. *Let* β_V^* *be the component of* β^* *corresponding to* V. *Then* $\beta_V^* \neq 0$.

Next, we investigate the role of random weighting, in particular how it affects the estimation, which motivates our algorithm. Consider a class of random weighting function

$$\mathcal{W}_B = \{w(x, B) > 0 : \mathbb{E}_{X \sim P_X}[w(X, B)|B] = 1, \forall B\},$$

where $B \sim P_B$ denotes the source of randomness in the weighting. $\mathcal{W}_B$ induces a class of covariate-shifted distributions $\mathcal{P} = \{P'(x, y) = P'(x)P(y|x) : P'(x) = w(x, B)P(x), w \in \mathcal{W}_B\}$. Note that each element of $\mathcal{P}$ is a distribution P' over (X, Y) that differs from the training distribution P only in the marginal distribution of X, because our weighting only depends on x.

We make the following assumptions on the weighting class $\mathcal{W}_\mathcal{B}$ and training distribution P and discuss how they reasonably suit the problem of covariate-shift with model misspeficiation.

Assumption 3.1. For all $P' \in \mathcal{P}$, the following two conditions hold: (i) $\mathbb{E}_{P'}[XX^\top]$ is constant; (ii) $\mathrm{Var}[\mathbb{E}_{P'}(Sg(S))] < \mathrm{Var}[\mathbb{E}_{P'}(Vg(S))]$.

Remark. Assumption (i) requires the distribution of covariates be expressive enough so that the random weighting will not affect its second order moments (although higher order moments can be generally affected). In Appendix, we show the existence of a sample weighting function such that the weighted distribution shares the same second-order moments with the original distribution under fairly general distributions of exponential families which have universal approximation capabilities (Sriperumbudur et al. 2017). Assumption (ii) indicates that the correlation between the invariant variable S and misspecified term $g(S)$ has lower variation than the correlation between the variant variable V and $g(S)$ across shifting covariate distributions P'. This assumption can be motivated from the invariance property of S and the unstable spurious correlation that varies with covariate shifts caused by V. We provide more discussion on the assumptions in Appendix and comment on the practical aspects at the end of Section 3.2 after we present the algorithm.

Given a random weighting $w(x, B)$ with $B \sim P_B$, let $\hat{\beta}_B = \arg\min_b \frac{1}{N} \sum_{i=1}^{N} [w(x_i, B)(y_i - b^\top x_i)^2]$ be the solution of weighted least squares. Let $\tilde{\beta}$ be the probability limit of $\hat{\beta}_B$ as $N \to \infty$ conditional on B. By studying $\tilde{\beta}$, we approximately integrate out the randomness from the finite sample while focus on the effect of random sample weighting B on the estimation.

Theorem 3.2. *Let* $\tilde{\beta}_S$ *and* $\tilde{\beta}_V$ *be the components of* $\tilde{\beta}$ *corresponding to* S *and* V *respectively. Under Assumption 3.1, we have* $\mathrm{Var}(\tilde{\beta}_S) < \mathrm{Var}(\tilde{\beta}_V)$.

See Appendix for the proof. Theorem 3.2 shows that as we conduct random sample weighting, $\tilde{\beta}_S$ exhibits a lower variance than $\tilde{\beta}_V$, indicating that the estimation of β_S is more stable and varies less than that of β_V. This provides a possibility to distinguish between S and V through the idea of invariant learning and therefore motivates us to propose the algorithm next.

3.2 Covariate-Shift Generalization via Random Sample Weighting

Now, we present covariate-shift generalization via random sample weighting (RSW), a simple non-parametric method that can produce sufficient heterogeneity effectively. Inspired by Theorem 3.2, through drawing random weighting functions $\{w(x, B_e)\}_{e=1}^E$ from a specific distribution $B_e \sim P_B$, we have $\mathrm{Var}(\tilde{\beta}_S) < \mathrm{Var}(\tilde{\beta}_V)$ across the distinct distributions induced by sample weights. That is to say

we can simulate heterogeneous environments via different weighting functions such that the estimation of then parameter associated with V tends to exhibit a larger variance across the simulated environments. On the other hand, Equation (6) suggests that the invariance constraint $\sum_{e\in\mathcal{E}_{tr}} \mathbb{D}(\theta, \Phi, e)$ would encourage the prediction model to rely on the variables with small variances of parameter estimation . Therefore, we can adopt the distinguishable property of invariance constraint to achieve stable prediction against covariate-shift that utilizes the invariant variables from the heterogeneous training environments produced by random sample weighting.

To employ the invariance constraint in linear models, referring to Liu et al. (2021b), we take a soft binary mask $M = [m_1, \ldots, m_d]$ (Yamada et al. 2020) as $\Phi(X)$, where $m_i = \max\{0, \min\{1, \mu_i + \epsilon\}\}$ is a clipped Gaussian variable parameterized by μ_i and ϵ is drawn from $\mathcal{N}(0, \sigma^2)$. Then we use the gradient norm penalty $\| \bigtriangledown_\theta \mathcal{L}^e(\theta) \odot M\|^2$ to measure the optimality of θ at each simulated environment, following IRM. As a result, after randomly sampling E weighting functions parameterized by $\{B_1, \ldots, B_E\} \sim P_B$, our goal is to minimize the following objective:

$$\mathcal{L} = 1/E \sum_{e=1}^{E} \mathcal{L}^e + \lambda \cdot \mathbb{D}(\theta, M, e) (= \| \bigtriangledown_\theta \mathcal{L}^e(\theta) \odot M\|^2)$$
$$\mathcal{L}^e = \sum_{i=1}^{N} w(x_i; B_e) \cdot \mathcal{L}(M \odot x_i, y_i; \theta) + \alpha \cdot \|M\|_0 \tag{7}$$

where $\|M\|_0 = \sum_{d=1}^{D} \mathrm{CDF}(\mu_d/\sigma)$ and CDF is the standard Gaussian CDF. Also, we can replace the invariance constraint in Equation (7) with $\|\mathrm{Var}(\bigtriangledown_\theta \mathcal{L}^e(\theta)) \odot M\|^2$ proposed by MIP (Koyama and Yamaguchi 2020). If the prediction model is nonlinear, one can directly optimize the invariant representation $\Phi(X)$ through minimizing $\mathbb{D}(\theta, \Phi, e) = \| \bigtriangledown_{\theta|\theta=1.0} \mathcal{L}^e(\theta \cdot \Phi)\|^2$ suggested by IRM, without using the mask M.

However, in practice with a finite number of sample weightings, it may happen that $\tilde{\beta}_V$ behaves more stably than $\tilde{\beta}_S$ due to randomness, which may mislead the invariant constraint. To this end, we propose RSW that randomly samples E new weighting functions at each optimization step of the prediction model with an auxiliary invariant constraint, which eventually approximates the infinite sample weightings as the algorithm proceeds.

We describe the algorithmic details in Algorithm 1. Since the complexity of calculating sample weights is only $\Omega(N * E)$, RSW simulates the heterogeneous environments for invariant learning without heavy costs. After the prediction model has meet enough environments, the invariance term in loss function will be guaranteed to have a positive effect.

Remark. Note that Assumption 3.1 that supports Theorem 3.2 constrains on the data distribution and the choice of sample weighting. However in the problem of covariate-shift generalization, we are mainly concerned about whether the learned prediction model discards the variant variables

Simulation 1: $p = 10, \lvert V_b\rvert = 2, Scale = 6$						
Method	$r = 2$			$r = 3$		
	Mean	Std	Worst	Mean	Std	Worst
ERM	0.5768	0.4420	1.1525	0.7122	0.5601	1.3393
ERM$^{\text{mixup}}$	0.5736	0.4387	1.1459	0.7022	0.5503	1.3191
STG	0.5765	0.4417	1.1522	0.7126	0.5605	1.3416
DWR	0.5167	0.3303	0.9195	0.6365	0.4640	1.1602
SRDO	0.5943	0.4587	1.1931	0.7038	0.5518	1.3245
DRO	0.5063	0.3025	0.9163	0.6212	0.2693	0.9402
JTT	0.5632	0.4255	1.1026	0.6992	0.5472	1.3124
EIIL	0.4763	0.0838	0.6437	0.5149	0.0978	0.7831
HRM	0.5041	0.0790	0.6656	0.5492	0.1168	0.8038
K$-$means	0.4906	0.2625	0.9701	0.4908	0.1992	0.8801
RSW$^{\text{I}}$	0.4234	**0.0749**	**0.5907**	0.4671	0.0929	0.6809
RSW$^{\text{M}}$	**0.4175**	0.1010	0.6226	**0.4653**	**0.0852**	**0.6512**
Simulation 2: $p = 10, \lvert V_b\rvert = 2, r = 2$						
Method	$Scale = 7$			$Scale = 8$		
	Mean	Std	Worst	Mean	Std	Worst
ERM	0.7179	0.5993	1.5055	0.7745	0.6389	1.6146
ERM$^{\text{mixup}}$	0.7145	0.5959	1.5008	0.7744	0.6389	1.6143
STG	0.7169	0.5981	1.5053	0.7747	0.6391	1.6135
DWR	0.6852	0.5596	1.4109	0.6355	0.4449	1.2148
SRDO	0.7051	0.5849	1.4711	0.7698	0.6311	1.5993
DRO	0.5341	0.2660	0.8772	0.6499	0.1648	0.8996
JTT	0.7235	0.6048	1.5183	0.7750	0.6398	1.6178
EIIL	0.4182	0.1578	0.6362	0.5563	0.0585	0.6518
HRM	0.4767	0.2983	1.0574	0.5513	0.1136	0.7325
K$-$means	0.4415	0.1957	0.7277	**0.5249**	0.0881	0.6696
RSW$^{\text{I}}$	**0.3849**	0.0929	0.5187	0.5338	0.0717	0.6670
RSW$^{\text{M}}$	0.3869	**0.0536**	**0.5134**	0.5391	**0.0690**	**0.6582**
Simulation 3: $\lvert V_b\rvert/\lvert V\rvert = 0.1, r = 0.2, Scale = 6$						
Method	$p = 20$			$p = 30$		
	Mean	Std	Worst	Mean	Std	Worst
ERM	0.5459	0.3911	1.0452	0.5883	0.4913	1.2709
ERM$^{\text{mixup}}$	0.5454	0.3909	1.0444	0.5890	0.4922	1.2722
STG	0.5463	0.3915	1.0457	0.5869	0.4901	1.2705
DWR	0.5910	0.4208	1.1330	0.5571	0.4591	1.196
SRDO	0.5467	0.3919	1.0461	0.5884	0.4914	1.2710
DRO	0.5537	0.3816	1.0323	0.5722	0.2693	1.0528
JTT	0.5491	0.3938	1.0510	0.5920	0.4945	1.2801
EIIL	0.5494	0.3936	1.0423	0.4147	0.1267	0.6365
HRM	0.5154	0.2721	0.9100	0.4303	0.1373	0.6645
K$-$means	0.5155	**0.1698**	0.8707	0.4375	0.1347	0.6812
RSW$^{\text{I}}$	0.4996	0.1721	0.8651	0.4158	**0.1281**	**0.6255**
RSW$^{\text{M}}$	**0.4978**	0.1728	**0.8680**	**0.4144**	0.1290	**0.6255**

Table 1: Results of experiments in synthetic data. To evaluate the performance of different benchmark models towards covariate-shift generalization, we conduct experiments under different settings, by varying the strengths of spurious correlations, the strengths of nonlinear term and the dimension of covariates, in simulation1, simulation2 and simulation3 respectively. We report the average (MEAN), STD and the worst-case of MSE in 10 test environments with distinct distributions. Compared to other baselines, RSW has the significant advantages in most of the settings.

V that carry unstable spurious correlations varying between training and test distributions, which tend to have larger variance of parameter. Hence, the practitioners may not have to elaborately design the weighting functions to meet the requirements, but get promising results with usual substances.

4 Experiment

In this section, we evaluate the effectiveness of proposed method RSW towards covariate-shift generalization, in comparison with the benchmark models. We carry out extensive experiments on both synthetic data and real-world datasets where the distribution shifts exist.

4.1 Baselines

We compare our proposed RSW with existing methods that are relevant to the problem we study. The baselines include ERM, ERM-mixup that pools data from different environments, STG (Yamada et al. 2020) (feature selection approach), stable learning methods (DWR (Kuang et al. 2020) and SRDO (Shen et al. 2020)), DRO (Sinha, Namkoong, and Duchi 2018) (Distributionary Robust Optimization), JTT (Liu et al. 2021a) (Just Train Twice), invariant learning using mixed data without domain labels (EIIL (Creager, Jacobsen, and Zemel 2021) and HRM (Liu et al. 2021b)), and invariant learning in created heterogeneous environments by K-means (MacQueen et al. 1967) on covariates. For fair comparison, all the methods are provided with training data from single source. See Appendix for the details about the implementation of baselines and the choice for weighting functions of RSW used in experiments.

4.2 Evaluation Metrics

For comprehensive evaluation of performance in the covariate-shift scenarios, we take the following three metrics. For convenience, we assume the model's accuracies of T test environments are $\{\text{acc}_t\}_{t=1}^T$. Then the three metrics are defined by:

- Average Accuracy ($\overline{\text{Acc}}$) $= \sum_{k=t}^T \text{acc}_t/T$.
- Standard Deviation (STD) of Accuracy (Acc_{std}) $= \left[\frac{1}{T-1}\sum_{t=1}^T(\text{acc}_t - \overline{\text{Acc}})^2\right]^{1/2}$.
- Worst-Case Accuracy (Acc_{worst}) $= \min_{t\in[T]} \text{acc}_t$.

4.3 Synthetic Data

Data Generation Process We consider the data generation mechanism that is called the confounder structure in literature of causality, where the relationships between invariant variables S, variant variables V and the outcome variable Y satisfy: $S \to Y \; and \; S \to V$.

Let $X = \{S_1, \ldots, S_{p_s}, V_1, \ldots, V_{p_v}\}$ be the observed covariates, where $p_s = p_v = p/2$ and p is the dimension of covariates. Hence, we can generate the observational data $\mathbf{X}$ with the help of an auxiliary quantity $\mathbf{Z}$ drawn from independent Gaussian distributions as following:

$$\mathbf{Z}_{,1}, \ldots, \mathbf{Z}_{,p_{s+1}} \overset{iid}{\sim} \mathcal{N}(0,1)$$
$$\mathbf{S}_{,i} = 0.8 * \mathbf{Z}_{,i} + 0.2 * \mathbf{Z}_{,i+1}$$
$$\mathbf{V}_{,j} = 0.8 * \mathbf{S}_{,j} + 0.2 * \mathbf{S}_{,j+1} + \epsilon_j (\sim \mathcal{N}(0, 0.3^2))$$

where $i, j = 1, 2, \cdots, p_s$, $\mathbf{S}_{,i}, \mathbf{V}_{,j}$ represent the values of S_i, V_j respectively, and we let $j+1 = (j+1) \bmod p_s$.

To elicit model misspecification, we introduce a nonlinear term in the form of multi-layer perceptron (MLP) into the generation function of outcome variable Y,

$$\begin{aligned} Y &= f(\mathrm{S}) + \mathcal{N}(0, 0.3^2) \\ &= [\beta_s, \beta_v]^T \cdot [S, V] + mlp(S_i, S_2, S_3; \psi) + \mathcal{N}(0, 0.3^2) \end{aligned}$$

where $\beta_s = \{\frac{1}{3}, -\frac{2}{3}, 1, -\frac{1}{3}, \frac{2}{3}, -1, \cdots\}$, $\beta_v = \vec{0}$, and the function mlp is a 3-Layer MLP (3x3x1) with parameters sampled from $\mathcal{U}(-Scale, Scale)$. By adjusting the value of $Scale$, we can change the strength of the nonlinear term, i.e., the degree of model misspecification for linear fitting.

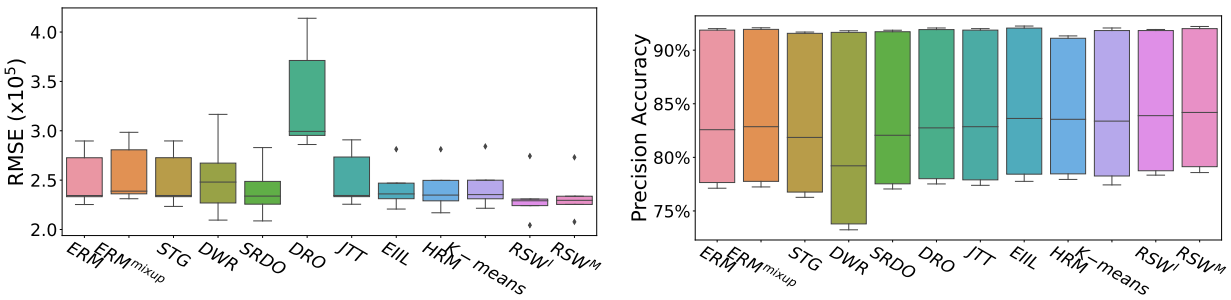
To test the generalization ability of models, it needs to generate a set of environments with distinct distributions of covariates $P(x)$ while keeping $P(y|x)$ invariant. Hence, we consider varying $P(v|s)$ across different environments, so that it brings about the spurious associations between V and Y. Specifically, we choose a subset of variant variables $V_b \subseteq V$ and perturb $P(v_b|s)$ via biased sample selection, leaving the others $V \setminus V_b$ still in the confounder structure. For each sample (x_i, y_i), the probability of it being selected is $\hat{P} = \prod_{V_b \subseteq V} g(s_i, v_{i,b}, r)$, where $g(\cdot)$ is a distance function and $v_{i,b}$ denotes the bth component of v_i. We apply two different functions in our experiments,

$$g_1(s_i, v_{i,b}, r) = |r|^{-5*|f(s_i) - sign(r)\cdot v_{i,b}|}$$
$$g_2(s_i, v_{i,b}, r) = PDF_{\mathcal{N}(sign(r)\cdot f(s_i),\ |r|^2)}\ (v_{i,b})$$

where $sign(r) = r/|r|$. Intuitively, the value of $|r|$ and $sign(r)$ control the strength and direction of spurious associations between V_b and Y. The larger value of $|r|$ implies the stronger correlations of V_b and Y in data using function $g_1(\cdot)$, while it is converse if $g_2(\cdot)$ is used. And $r > 0$ means positive correlations and vice versa. As a result, we can employ r to define different environments.

Experimental Settings

- **Simulation 1** We use $g_1(\cdot)$ for biased sample selection and set $Scale = 6$. First, we collect 200 test samples from each of 10 environments where $r \in \{\pm 3.5, \pm 3, \pm 2.5, \pm 2, \pm 1.5\}$ (totally 2000 test samples). Then we optimize the models using the 1000 training samples collected from a single environment where $r = 2/3$.
- **Simulation 2** The data generation and collection is similar to simulation1, except that we fix the $r = 2$ in training environment, and modify the $Scale \in \{7, 8\}$ to vary the strength of nonlinear term.



(a) House Price Prediction

(b) People Income Prediction

Figure 1: Results of experiments in real-world datasets. We report the performance of algorithms in all the test environments by boxplot. From the state of results' dispersion, it is apparent to see the advantage of RSW, specially in term of the worst-case performance.

- **Simulation 3** We use $g_2(\cdot)$ for biased sample selection and set $Scale = 6$. The 2000 test samples are uniformly collected from 10 environments where $r \in \{\pm 1, \pm 0.8, \pm 0.4, \pm 0.2, \pm 0.1\}$, and 1000 training samples are collected from a single environment where $r = 0.2$. We fix $|V_b|/V = 0.1$ and change the dimension of observed variables X in experiment.

4.4 Experimental Results

From the results in Table 1, we can observe that:

1. The ERM is easily affected by spurious associations, in terms of its unstable performance (STD) and poor performance in the worst case (Worst). The influence is much more serious when increasing the strength of selection bias and model misspecification. Mixing up multi-environments can bring a little benefit, but the mixture of these environments is still biased.
2. The STG fails to pick up the true variables through the feature selection approaches from shifted distributions.
3. Stable learning methods (DWR and SRDO) work if they can achieve the weights that indeed make covariates independent. But in some scenarios where it is difficult to calculate the weights from finite samples, e.g. the high dimensional variables, they may bring some negative effects.
4. The methods (DRO and JTT) that focus on optimizing the prediction of samples with larger empirical errors alleviate the gap of different sub-populations. But it depends on the support of a distribution (such as DRO suffers from over-pessimism problem encountering large distribution set), resulting in their unstable performances across different settings.
5. The methods (HRM and EIIL) raise the generalization ability through the invariance constraint, but they still fails in some cases. Although the invariant relationship can help to identify the heterogeneity, it would introduce the model misspecification of predicting the outcome variable into the partition of environments, and mislead the iterative learning process eventually.
6. The K-means can cluster the data into quite valuable environments sometimes, but not at other times, relying on the heterogeneity of sub-populations inside training distribution, and the initial clustering centers.
7. Compared to other baselines, all the environments simulated by random sample weighting can guarantee to have sufficient heterogeneity in expectation to encourage a stable prediction model utilizing invariant correlations. Hence, we achieve the best performance in most of the cases.

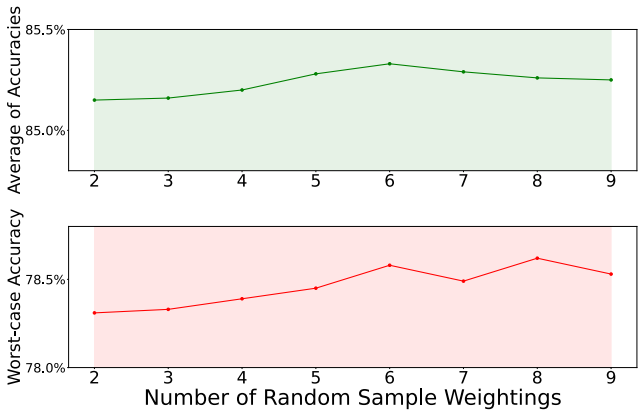


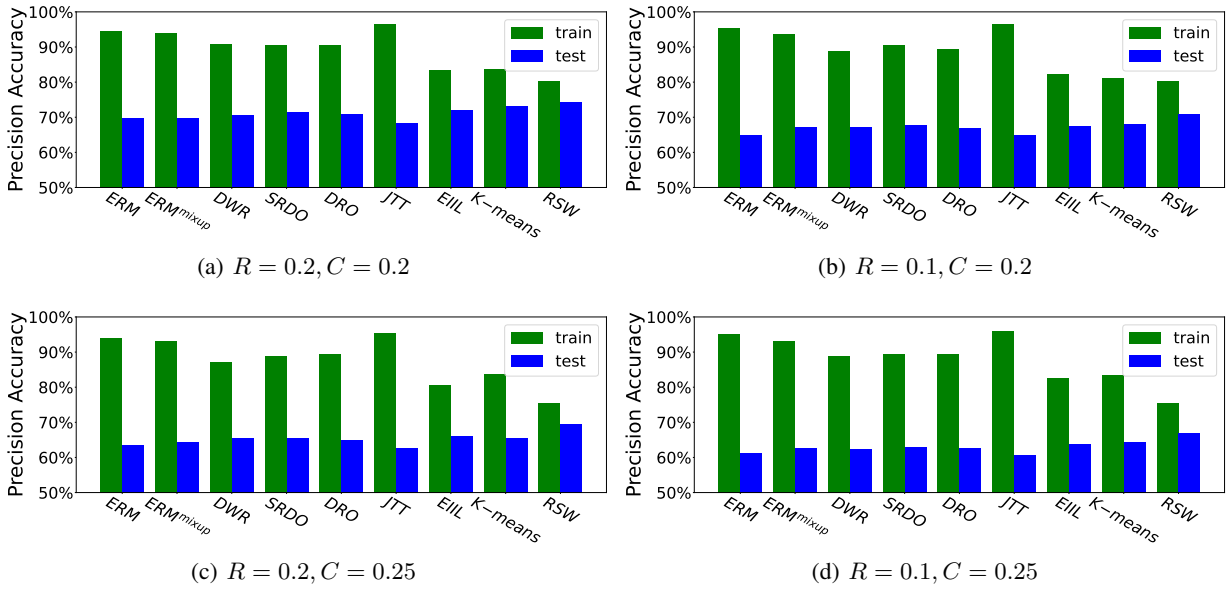
Figure 2: Parameter analysis on the number of sample weightings at each optimization step. With sufficient heterogeneity provided, the performance curve finally converges.

4.5 Real-World Data

Due to the space limit, please see Appendix for the detailed descriptions of real datasets used here.

House Price Prediction In this experiment, we use a real-world regression dataset[1] (Kaggle) of house sales prices from King County, USA. Specifically, we split the dataset into 6 periods (each covers a time span of two decades) between 1900∼2015 according to *the built year of house*. To

[1] https://www.kaggle.com/datasets/harlfoxem/housesalesprediction



(a) $R = 0.2, C = 0.2$ (b) $R = 0.1, C = 0.2$

(c) $R = 0.2, C = 0.25$ (d) $R = 0.1, C = 0.25$

Figure 3: Results of experiments in CS-Colored MNIST dataset. Compared to other baselines, RSW maximizes the use of digital information in training phase (less misled by color), leading to the best performance in test data where $R = 0.9$.

test the stability, we train all the methods on the first period ($[1900, 1919]$) and test if they can predict the transaction price on the other periods respectively. From the results (Figure 1(a)), we observe that: 1) for the high-dimensional and sparse tabular data, some methods even perform worse than ERM due to the violated assumption or difficulty in tuning, while ours is not disturbed by that; 2) our method has a more prominent advantage in the worst case, which verifies its effectiveness.

People Income Prediction In this experiment, we use the Adult dataset[2] (Kohavi 1996), of which the task is to predict whether the personal income exceeds $50K/yr$ based on census data. We split the dataset into 10 environments according to the combination of attributes race and sex. Like in House Price Prediction, we train all the methods on the first environment $(White, Female)$ and test them on the other environments respectively. The results (Figure 1(b)) show the superiority of our method. Also, we conduct the hyper-parameter analysis for the number of sample weighting at each optimization step of prediction model. In Figure 2, it is apparent that the performance of RSW gradually converges as the number increases, which implies weighting samples for finite times can already produce sufficient heterogeneity.

CS-Colored MNIST In this experiment, we use the CS-Colored MNIST dataset (Ahuja et al. 2021) that simulates the covariate-shift in image data. It assigns the digits to 2 categories, and colors the digits with either red or green in ratio R and $(1 - R)$ respectively according to the labels, resulting in the spurious correlations. Then label flipping happens with a probability of C to make classifier easily rely on color variable. Here, we set $R = 0.1$ or $R = 0.2$ in training data ($R = 0.9$ in test data), and vary $C = 0.2$ or $C = 0.25$. A 3-Layer MLP is taken as prediction model for image, which is insufficient to capture the image patterns, leading to model misspecification.

We report the test results of benchmark models having competitive performances in Figure 3. The coloring ratio R misleads the classifier to make prediction based on color information but not the original image, causing its poor performance if the spurious association of color and label is reverse. Through perturbing the variant relationships using ceaseless random sample weights, with access to auxiliary invariance constraint, the classifier is enforced to rely on the color as little as possible when minimizing the empirical errors (the drop in training accuracy), while obtaining the generalization ability to the test distribution (the improvement in test accuracy).

5 Conclusion

In this paper, we study the covariate-shift generalization problem under model misspecification. We theoretically show that under appropriate conditions, random sample weighting can produce sufficient heterogeneity in the sense of leading to different variances in estimating the parameters associated with invariant and variant covariates, which can be exploited by common invariance constraints in finding the invariant variables for stable prediction under covariate shifts. Motivated by this result, we propose the random sample weighting (RSW) algorithm to simulate heterogeneous environments and encourage the prediction model to get rid of unstable correlations via invariance constraints. The extensive experimental results well support our claims and demonstrate the advantages of our proposal.

[2]https://archive.ics.uci.edu/ml/datasets/adult

Acknowledgements

This work was supported in part by National Key R&D Program of China (No. 2018AAA0102004), National Natural Science Foundation of China (No. U1936219, 62141607), Beijing Academy of Artificial Intelligence (BAAI).

References

Ahuja, K.; Wang, J.; Dhurandhar, A.; Shanmugam, K.; and Varshney, K. R. 2021. Empirical or Invariant Risk Minimization? A Sample Complexity Perspective. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization. *CoRR*, abs/1907.02893.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Mach. Learn.*, 79(1-2): 151–175.

Creager, E.; Jacobsen, J.; and Zemel, R. S. 2021. Environment Inference for Invariant Learning. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 2189–2200. PMLR.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. S. 2017. Domain-Adversarial Training of Neural Networks. In Csurka, G., ed., *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, 189–209. Springer.

Ghifary, M.; Kleijn, W. B.; Zhang, M.; and Balduzzi, D. 2015. Domain Generalization for Object Recognition with Multi-task Autoencoders. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2551–2559. IEEE Computer Society.

Hassanpour, N.; and Greiner, R. 2019. CounterFactual Regression with Importance Sampling Weights. In Kraus, S., ed., *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 5880–5887. ijcai.org.

Koh, P. W.; Sagawa, S.; Marklund, H.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Gao, I.; Lee, T.; David, E.; Stavness, I.; Guo, W.; Earnshaw, B.; Haque, I.; Beery, S. M.; Leskovec, J.; Kundaje, A.; Pierson, E.; Levine, S.; Finn, C.; and Liang, P. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 5637–5664. PMLR.

Kohavi, R. 1996. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In Simoudis, E.; Han, J.; and Fayyad, U. M., eds., *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, 202–207. AAAI Press.

Koyama, M.; and Yamaguchi, S. 2020. Out-of-Distribution Generalization with Maximal Invariant Predictor. *CoRR*, abs/2008.01883.

Krueger, D.; Caballero, E.; Jacobsen, J.; Zhang, A.; Binas, J.; Zhang, D.; Priol, R. L.; and Courville, A. C. 2021. Out-of-Distribution Generalization via Risk Extrapolation (REx). In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 5815–5826. PMLR.

Kuang, K.; Xiong, R.; Cui, P.; Athey, S.; and Li, B. 2020. 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*, 4485–4492. AAAI Press.

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

Lee, W.; Kim, H.; and Lee, J. 2021. Compact class-conditional domain invariant learning for multi-class domain adaptation. *Pattern Recognit.*, 112: 107763.

Li, D.; Yang, Y.; Song, Y.; and Hospedales, T. M. 2017. Deeper, Broader and Artier Domain Generalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 5543–5551. IEEE Computer Society.

Liu, E. Z.; Haghgoo, B.; Chen, A. S.; Raghunathan, A.; Koh, P. W.; Sagawa, S.; Liang, P.; and Finn, C. 2021a. Just Train Twice: Improving Group Robustness without Training Group Information. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 6781–6792. PMLR.

Liu, J.; Hu, Z.; Cui, P.; Li, B.; and Shen, Z. 2021b. Heterogeneous Risk Minimization. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 6804–6814. PMLR.

MacQueen, J.; et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, 281–297. Oakland, CA, USA.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain Generalization via Invariant Feature Representation. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, 10–18. JMLR.org.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70: 41–55.

Rudin, C.; and Ustun, B. 2018. Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice. *Interfaces*, 48(5): 449–466.

Shen, Z.; Cui, P.; Zhang, T.; and Kuang, K. 2020. Stable Learning via Sample Reweighting. 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*, 5692–5699. AAAI Press.

Sinha, A.; Namkoong, H.; and Duchi, J. C. 2018. Certifying Some Distributional Robustness with Principled Adversarial Training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Sriperumbudur, B.; Fukumizu, K.; Gretton, A.; Hyvärinen, A.; and Kumar, R. 2017. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18.

Wong, A.; Hryniowski, A.; and Wang, X. Y. 2020. Insights into Fairness through Trust: Multi-scale Trust Quantification for Financial Deep Learning. *CoRR*, abs/2011.01961.

Xu, R.; Zhang, X.; Cui, P.; Li, B.; Shen, Z.; and Xu, J. 2022a. Regulatory Instruments for Fair Personalized Pricing. In Laforest, F.; Troncy, R.; Simperl, E.; Agarwal, D.; Gionis, A.; Herman, I.; and Médini, L., eds., *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, 4–15. ACM.

Xu, R.; Zhang, X.; Shen, Z.; Zhang, T.; and Cui, P. 2022b. A Theoretical Analysis on Independence-driven Importance Weighting for Covariate-shift Generalization. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 24803–24829. PMLR.

Yamada, Y.; Lindenbaum, O.; Negahban, S.; and Kluger, Y. 2020. 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*, 10648–10659. PMLR.

Zhang, X.; Cui, P.; Xu, R.; Zhou, L.; He, Y.; and Shen, Z. 2021. Deep Stable Learning for Out-of-Distribution Generalization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 5372–5382. Computer Vision Foundation / IEEE.