Causal Balancing for Domain Generalization

Xinyi Wang [1] **Michael Saxon** [1] **Jiachen Li** [1] **Hongyang Zhang** [2] **Kun Zhang** [3] **William Yang Wang** [1]

Abstract

While machine learning models rapidly advance the state-of-the-art on various real-world tasks, out-of-domain (OOD) generalization remains a challenging problem given the vulnerability of these models to spurious correlations. We propose a causally-motivated balanced mini-batch sampling strategy to train robust classifiers that is minimax optimal across a diverse enough environment space, by utilizing multiple training sets from different environments. We provide an identifiability guarantee of the latent covariates in the proposed causal graph and show that our proposed approach samples train data from a balanced, spurious-free distribution under an ideal scenario. Experiments are conducted on three domain generalization datasets, demonstrating empirically that our balanced mini-batch sampling strategy improves the performance of four different established domain generalization model baselines compared to the random mini-batch sampling strategy.

1. Introduction

Machine learning is achieving tremendous success in many fields with useful real-world applications (Silver et al., 2016; Devlin et al., 2019; Jumper et al., 2021). However, machine learning models often fail to generalize to out-of-domain (OOD) data sampled from unseen environments (Quiñonero-Candela et al., 2009; Szegedy et al., 2014). One explanation for such a failure is that the models are prone to learning spurious correlations that change between environments.

Recently, various causal-inspired (Pearl, 2009) methods have been proposed to improve the OOD generalizability by considering the invariance of causal features or the underlying causal mechanism through which data is generated. Such methods often aim to find invariant data representations using new loss function designs that incorporate the invariance conditions across different domains into the training process (Arjovsky et al., 2020; Mahajan et al., 2021; Liu et al., 2021; Lu et al., 2022; Wald et al., 2021). Unfortunately, these approaches have to contend with trade-offs between weak linear models or approaches without theoretical guarantees (Arjovsky et al., 2020; Wald et al., 2021), and empirical studies have shown their utility in the real world to be questionable (Gulrajani & Lopez-Paz, 2020).

In this paper, we first demonstrate that the Bayes optimal classifier trained on a balanced (spurious-free) distribution is minmax optimal across all environments. Then we propose a two-step method to create such balanced distribution from multiple train datasets collected form different environments: (1) learn the observed data distribution using a variational autoencoder (VAE) by *latent covariate learning*, and (2) use the learned latent covariate to create *balanced mini-batches* that follow a balanced distribution. Because the only modification at train time is a resampling of train examples, it makes our method lightweight and highly flexible, enabling seamless incorporation with off-the-shelf domain generalization methods (Gulrajani & Lopez-Paz, 2020; Sagawa et al., 2019; Sun & Saenko, 2016).

Our contributions are as follows: (1) We propose the balanced distribution without spurious correlation and prove that it can produce mimmax optimal classifiers for OOD generalization; (2) We demonstrate that the source of spurious correlation, as a latent variable, can be identified given a large enough set of training environments under mild conditions in a nonlinear setting; (3) We propose a novel balanced mini-batch sampling algorithm that, in an ideal scenario with exact matches of the true source of spurious correlation, can remove the spurious correlations in the observed data distribution; (4) Our empirical results show that our two-phased method obtains significant performance gain on three domain generalization datasets, ColoredMNIST (Arjovsky et al., 2020), PACS (Li et al., 2017) and TerraIncognita (Beery et al., 2018), across four different domain generalization methods over a random sampling strategy.

[1]Department of Computer Science, University of California, Santa Barbara, USA [2]David R. Cheriton School of Computer Science, University of Waterloo, Canada [3]Department of Philosophy, Carnegie Mellon University, USA. Correspondence to: Xinyi Wang <xinyi_wang@ucsb.edu>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2022 by the author(s).

(a) Observed distribution $p(X, Y|E = e)$

(b) Balanced distribution $\hat{p}(X, Y|E = e)$

Figure 1. The causal graphical model assumed for data generation process in environment $e \in \mathcal{E}$. Shaded nodes means being observed and white nodes means not being observed. Black arrows means causal relations invariant across different environments. Red dashed line means correlation varies across different environments.

2. Preliminaries

2.1. Problem Setting

We consider a standard domain generalization setting with a potentially high-dimensional variable X (e.g. an image), a label variable Y and a discrete environment (or domain) variable E in the sample spaces $\mathcal{X}, \mathcal{Y}, \mathcal{E}$, respectively. Here we focus on the classification problems with $\mathcal{Y} = \{1, 2, ..., m\}$ and $\mathcal{X} \in \mathbb{R}^d$. We assume that the training data are collected from a finite subset of training environments $\mathcal{E}_{\text{train}} \subset \mathcal{E}$. The training data $\mathcal{D}^e = \{(x_i^e, y_i^e)\}_{i=1}^{N^e}$ is then sampled from the distribution $p^e(X, Y) = p(X, Y|E = e)$ for all $e \in \mathcal{E}_{\text{train}}$. Our goal is to learn a classifier $C_\psi : \mathcal{X} \to \mathcal{Y}$ that performs well in a new, unseen environment $e_{test} \notin \mathcal{E}_{\text{train}}$.

We assume that there is a data generation process of the observed data distribution $p^e(X, Y)$ represented by an underlying structural causal model (SCM) shown in Figure 1a. More specifically, we assume that X is caused by label Y, an unobserved latent variable Z (with sample space $\mathcal{Z} \in \mathbb{R}^n$) and an independent noise variable ϵ with the following formulation:

$$X = \mathbf{f}(Y, Z) + \epsilon = \mathbf{f}_Y(Z) + \epsilon \tag{1}$$

Here, we assume the causal mechanism is invariant across all environments $e \in \mathcal{E}$ and we further characterize $\mathbf{f}$ with the following assumption:

Assumption 2.1. $\mathbf{f} : \{1, 2, ..., m\} \times \mathcal{Z} \to \mathcal{X}$ is injective. $\mathbf{f}^{-1} : \mathcal{X} \to \{1, 2, ..., m\} \times \mathcal{Z}$ is the left inverse of $\mathbf{f}$.

Note that this assumption forces the generation process of X to consider both Z and Y instead of only one of them. Suppose ϵ has a known probability density function $p_\epsilon > 0$. Then we have

$$p_{\mathbf{f}}(X|Z, Y) = p_\epsilon(X - \mathbf{f}_Y(Z)) \tag{2}$$

While the causal mechanism is invariant across environments, we assume that the correlation between label Y and latent Z is environment-variant and Z should exclude Y information. i.e., Y cannot be recovered as a function of Z. If Y is a function of Z, the generation process of X can completely ignore Y and f would not be injective. Z can be understood as a potentially massive set of latent features that do not determine the "Y-ness" and that can be spuriously correlated with environment E.

Let $e \in \mathcal{E}$ index a family of distributions $\mathcal{F} = \{p^e(X, Y, Z) = p_{\mathbf{f}}(X|Z, Y)p^e(Z|Y)p^e(Y)\}_e$, where $p^e(Z|Y) > 0$ and $p^e(Y) > 0$. Note that any mixture of distributions from $\mathcal{F}$ would also be a member of $\mathcal{F}$.

In this setting, we can see that the correlation between X and Y would vary for different values of e. We argue that the correlation $Y \leftrightarrow Z \to X$ is not stable in an unseen environment $e \notin \mathcal{E}_{\text{train}}$ as it involves E and we only want to learn the stable causal relation $Y \to X$. However, it is inevitable that the learned predictor may absorb the unstable relation between X and Y if we simply train it on the observed train distribution $p^e(X, Y)$ with empirical risk minimization. The causal graphs of two examples of the realization of our data generation model are shown in Appendix A.

2.2. Balanced Distribution

To avoid learning the unstable relations, we propose to consider a balanced distribution $p^B(X, Y, Z)$ such that $Y \perp\!\!\!\perp_B Z$ while the causal mechanism $Z \to X \leftarrow Y$ unchanged, as shown in Figure 1b, which is defined below:

Definition 2.2. A **balanced distribution** can be written as $p^B(X, Y, Z) = p_{\mathbf{f}}(X|Y, Z)p^B(Z)p^B(Y)$, where $p^B(Y) = \mathrm{U}\{1, 2, ..., m\}$ and $Y \perp\!\!\!\perp_B Z$.

In this new distribution, X and Y is only correlated through the stable causal relation $Y \to X$. Here we do not specify $p^B(Z)$. Note that $p^B(X|Y, Z) = p_{\mathbf{f}}(X|Y, Z)$ is a result of the unchanged causal mechanism $Z \to X \leftarrow Y$, and that $p^B(X, Y, X) \in \mathcal{F}$ can also be regarded as from an environment $B \in \mathcal{E}$. Under an additional conditional independence assumption $Y \perp\!\!\!\perp_B Z|X$, we can prove that $p^B(Y|X)$ is invariant for any choice of $p^B(Z)$. Then we have the following theorem[1]:

Theorem 2.3. *Consider a classifier $C_\psi(X) = \operatorname{argmax}_Y p_\psi(Y|X)$ with parameter ψ. We denote the cross entropy loss of such a classifier on environment e by $L^e(p_\psi(Y|X)) = -\mathbb{E}_{p^e(X,Y)} \log p_\psi(Y|X)$. Assume that (1) $Y \perp\!\!\!\perp_B Z|X$, and (2) $\mathcal{E}$ satisfies:*

$$\forall e \in \mathcal{E}, Y \not\perp\!\!\!\perp_{p^e} Z \implies \exists e' \in \mathcal{E} \ s.t. \\ L^{e'}(p^e(Y|X)) - L^{e'}(p^B(Y)) > 0 \tag{3}$$

*Then the Bayes optimal classifier trained on any balanced distribution $p^B(X, Y)$ is a **minimax optimal classifier** with respect to cross entropy loss across all environments in $\mathcal{E}$:*

$$p^B(Y|X) = \operatorname*{argmin}_{p_\psi \in \mathcal{F}} \max_{e \in \mathcal{E}} L^e(p_\psi(Y|X)) \tag{4}$$

[1] See Appendix B for all the proofs.

The first assumption implies the noise variable ϵ can be disentangled into (ϵ_Y, ϵ_Z), such that there exist functions $\mathbf{g}_Y, \mathbf{g}_Z$ with $(Y, Z) = \mathbf{f}^{-1}(X - \epsilon) = (\mathbf{g}_Y(X - \epsilon_Y), \mathbf{g}_Z(X - \epsilon_Z))$. The second assumption implies that the environment space $\mathcal{E}$ is large and diverse enough such that a perfect classifier on one environment will always perform worse than random guessing on some other environment. Under these two assumptions, no other Byes optimal classifier produced by an environment in $\mathcal{E}$ would have a better worst case OOD performance than the balanced distribution.

3. Method

We propose a two-phased method that first use an VAE to learn the underlying data distribution $p^e(X, Y, Z)$ with latent covariate Z for each $e \in \mathcal{E}_{\text{train}}$, and then use the learned distribution to calculate a balancing score to create a balanced distribution based on the training data.

3.1. Latent Covariate Learning

We argue that the underlying joint distribution of $p^e(X, Y, Z)$ can be learned and identified by a VAE, given a sufficiently large set of train environments $\mathcal{E}_{\text{train}}$.

To specify the correlation between Z and Y, we assume that the conditional distribution $p^e(Z|Y)$ is conditional factorial with an exponential family distribution:

Assumption 3.1. The correlation between Y and Z in environment e is characterized by $p^e_{T,\lambda}(Z|Y)$ as follows:

$$p^e_{T,\lambda}(Z|Y) = \prod_{i=1}^{n} \frac{Q_i(Z_i)}{W^e_i(Y)} \exp\Big[\sum_{j=1}^{k} T_{ij}(Z_i)\lambda^e_{ij}(Y)\Big] \tag{5}$$

where Z_i is the i-th element of Z, $\mathbf{Q} = [Q_i]_i : \mathcal{Z} \to \mathbb{R}^n$ is the base measure, $\mathbf{W}^e = [W^e_i]_i : \mathcal{Y} \to \mathbb{R}^n$ is the normalizing constant, $\mathbf{T} = [T_{ij}]_{ij} : \mathcal{Z} \to \mathbb{R}^{nk}$ is the sufficient statistics, and $\lambda^e = [\lambda^e_{ij}]_{ij} : \mathcal{Y} \to \mathbb{R}^{nk}$ are the Y dependent parameters.

Here n is the dimension of the latent variable Z, and k is the dimension of each sufficient statistic determined by the type of chosen exponential family distribution. The simplified conditional factorial prior assumption is from the mean-field approximation, which can be expressed as a closed form of the true prior (Blei et al., 2017). Note that the exponential family assumption is not very restrictive as it has universal approximation capabilities (Sriperumbudur et al., 2017).

We then consider the following conditional generative model in each environment $e \in \mathcal{E}_{\text{train}}$, with parameters $\theta = (\mathbf{f}, \mathbf{T}, \lambda)$:

$$p^e_\theta(X, Z|Y) = p_{\mathbf{f}}(X|Z, Y)p^e_{\mathbf{T},\lambda}(Z|Y) \tag{6}$$

We use a VAE to estimate the above generative model with a variational approximation $q^e_\phi(Z|X, Y)$ of the prior probability of latent variable $p^e_\theta(Z|X, Y)$. We denote the empirical data distribution given by dataset $\mathcal{D}^e = \{(x^e_i, y^e_i)\}_{i=1}^{N^e}$ collected from environment e. The evidence lower bound (ELBO) of the data log-likelihood in each environment $e \in \mathcal{E}_{\text{train}}$ is then defined as follows:

$$\begin{aligned}&\mathbb{E}_{q_{\mathcal{D}^e}}\left[\log p^e_\theta(X|Y)\right] \geq \mathcal{L}^e(\theta, \phi) := \\ &\mathbb{E}_{q_{\mathcal{D}^e}}\Big[\mathbb{E}_{q^e_\phi(Z|X,Y)}\left[\log p_\theta(X|Z, Y)\right] \\ &- KL(q^e_\phi(Z|X, Y)||p^e_\theta(Z|Y))\Big]\end{aligned} \tag{7}$$

The KL-divergence term can be calculated analytically. To sample from the variational distribution $q^e_\phi(Z|X, Y)$, we use reparameterization trick (Kingma & Welling, 2013).

We then maximize the above ELBO $\frac{1}{|\mathcal{E}_{\text{train}}|}\sum_{e\in\mathcal{E}_{\text{train}}} \mathcal{L}^e(\theta, \phi)$ over all training environments to obtain model parameters (θ, ϕ). To show that we can uniquely recover the latent variable Z up to some simple transformations, we want to show that the model parameter θ is identifiable up to some simple transformations. That is, for any $\{\theta = (\mathbf{f}, \mathbf{T}, \lambda), \theta' = (\mathbf{f}', \mathbf{T}', \lambda')\} \in \Theta$,

$$p^e_\theta(X|Y) = p^e_{\theta'}(X|Y), \forall e \in \mathcal{E}_{\text{train}} \implies \theta \sim \theta' \tag{8}$$

where Θ is the parameter space and $\sim$ represents an equivalent relation. Specifically, we consider the following equivalence relation from (Motiian et al., 2017):

Definition 3.2. If $(\mathbf{f}, \mathbf{T}, \lambda) \sim_A (\mathbf{f}', \mathbf{T}', \lambda')$, then there exists an invertible matrix $A \in \mathbb{R}^{nk\times nk}$ and a vector $\mathbf{c} \in \mathbb{R}^{nk}$, such that $\mathbf{T}(\mathbf{f}^{-1}(x)) = A\mathbf{T}'(\mathbf{f}'^{-1}(x)) + \mathbf{c}, \forall x \in \mathcal{X}$.

When the underlying model parameter θ^* can be recovered by perfectly fitting the data distribution $p^e_{\theta^*}(X|Y)$ for all $e \in \mathcal{E}_{\text{train}}$, the joint distribution $p^e_{\theta^*}(X, Z|Y)$ is also recovered. This further implies the recovery of the prior $p^e_{\theta^*}(Z|Y)$ and the true latent variable Z^*.

The identifiablity of our proposed latent covariate learning model can then be summarized as follows:

Theorem 3.3. *Suppose we observe data sampled from the generative model defined according to Equation (6), with parameters $\theta = (\mathbf{f}, \mathbf{T}, \lambda)$. In addition to Assumption 2.1 and Assumption 3.1, we assume the following conditions holds: (1) The set $\{x \in \mathcal{X}|\phi_\epsilon(x) = 0\}$ has measure zero, where ϕ_ϵ is the characteristic function of the density p_ϵ. (2) The sufficient statistics T_{ij} are differentiable almost everywhere, and $(T_{ij})_{1\leq j\leq k}$ are linearly independent on any subset of $\mathcal{X}$ of measure greater than zero. (3) There exist $nk + 1$ distinct points $(y_0, e_0), \ldots, (y_{nk}, e_{nk})$ such that the $nk \times nk$ matrix*

$$\mathbf{L} = (\lambda^{e_1}(y_1) - \lambda^{e_0}(y_0), \ldots, \lambda^{e_{nk}}(y_{nk}) - \lambda^{e_0}(y_0)) \tag{9}$$

*is invertible. Then we have the parameters $\theta = (\mathbf{f}, \mathbf{T}, \lambda)$ are $\sim_A$-**identifiable**.*

Note that the last assumption in Theorem 3.3 implies that the product space $\mathcal{Y} \times \mathcal{E}_{\text{train}}$ has to be large enough to ensure the identifiability of θ when perfectly fitting the given

training data distribution. i.e. We need $m|\mathcal{E}_{\text{train}}| > nk$. The invertibility of $\mathbf{L}$ implies that $\lambda^{e_i}(y_i) - \lambda^{e_0}(y_0)$ need to be orthogonal to each other which further implies the diversity of environment space $\mathcal{E}$.

3.2. Balanced mini-batch sampling

We consider a classic method that has been widely used in the average treatment effect (ATE) estimation — balancing score matching — to sample balanced mini-batches that mimic a balanced distribution shown in Figure 1b.

Causal effect estimation studies the effect a treatment would have had on a unit which in reality received another treatment. A causal graph similar to Figure 1a is usually considered in a causal effect estimation problem, where Z is called the covariate (e.g. a patient profile), which is observed before treatment $Y \in \{0, 1\}$ (e.g. taking drug or placebo) is applied. We denote the effect of receiving a specific treatment $Y = y$ as X_y. Note that this causal graph implies that we make the Strong Ignorability assumption. i.e. Z includes all variables that are related to both X and Y.

In the case of a binary treatment, the ATE is defined as the expected difference of effect after receiving different treatments: $\mathbb{E}[X_1 - X_0]$ (e.g. difference in blood pressure). For a randomized controlled trial, we can directly estimate the difference between $\mathbb{E}[X|Y = 1]$ and $\mathbb{E}[X|Y = 0]$ from the observed data as the true treatment effect, as in this case we force $Z \perp\!\!\!\perp Y$ and there would not be systematical difference between units exposed to one treatments and units exposed to another.

However, in most observed datasets, Z is correlated with Y. Thus $\mathbb{E}[X|Y = 1]$ and $\mathbb{E}[X|Y = 0]$ are not directly comparable. We can then use balancing score $b(Z)$ (Dawid, 1979) to de-correlate Z and Y:

Definition 3.4. A **balancing score** $b(Z)$ is a function of covariate Z s.t. $Z \perp\!\!\!\perp Y|b(Z)$.

The ATE can then be estimated by matching units with same balancing score but different treatments: $\mathbb{E}_{b(Z)}\left[\mathbb{E}[X|Y = 1, b(Z)] - \mathbb{E}[X|Y = 0, b(Z)]\right]$. There is a wide range of functions of Z that can be used as a balancing score, where the scalar propensity score $p(Y = 1|Z)$ is the coarsest one and the covariate Z itself is the finest one (Rosenbaum & Rubin, 1983). To extend this statement to non-binary treatments, we first define propensity score $s(Z)$ for $Y \in \mathcal{Y} = \{1, 2, ..., m\}$ as a vector:

Definition 3.5. The **propensity score** for $Y \in \{1, 2, ..., m\}$ is $s(Z) = [p(Y = y|Z)]_{y=1}^{m}$.

We then have the following theorem that applies to the vector version of propensity score $s(Z)$:

Theorem 3.6. *Let $b(Z)$ be a function of Z. Then $b(Z)$ is a balancing score, if and only if $b(Z)$ is finer than $s(Z)$. i.e. exists a function g such that $s(Z) = g(b(Z))$.*

We use $b^e(Z)$ to denote the balancing score for a specific environment e. The propensity score in a training environment e would then be $s^e(Z) = [p_\theta^e(Y = y|Z)]_{y=1}^{m}$, which can be derived from the learned conditional prior $p_\theta^e(Z|Y)$:

$$p_\theta^e(Y = y|Z) = \frac{p_\theta^e(Z|Y = y)p^e(Y = y)}{\sum_{i=1}^{m} p_\theta^e(Z|Y = i)p^e(Y = i)} \tag{10}$$

where $p^e(Y = i)$ can be directly estimated from the training data $\mathcal{D}^e$.

We propose to construct balanced mini-batches by matching $1 \leq a \leq m - 1$ examples with a different label Y but the same/closest balancing score $b^e(Z)$ for each example sampled from the training environment e. (The detailed sampling algorithm is shown in Appendix A.)

With perfect match at every step (i.e., $b^e(z_j) = b^e(z)$) and $a = m - 1$, we can obtain a completely balanced mini-batch sampled from the balanced distribution with $Y \perp\!\!\!\perp Z$. However, an exact match of balancing score is unlikely in reality, so the quality of matched data point would likely be lower than the referencing data point in terms of having the same balancing score. This can be mitigated by choosing a smaller a. However, this would make Y and Z not completely independent. In fact, if we have exact match of balancing score, the larger a is, the weaker the correlation between Y and Z would be. So in practice, the choice of a reflects a trade-off between the balancing score matching quality and the degree of dependency between Y and Z. The above arguments can be summarized as below:

Theorem 3.7. *A balanced mini-batch with exact matches of balancing score and $a = m - 1$ can be regarded as sampling from the balanced distribution over all training environments $p^B(X, Y, Z)$. In general, the balanced mini-batch can be regarded as sampling from a semi-balanced distribution with $\hat{p}^B(Y|Z, E) = \frac{1}{a+1}(\frac{a}{m-1} + \frac{m-a-1}{m-1}p(Y|Z, E))$.*

4. Experiments

To verify the effectiveness of our proposed mini-batch balancing method, we conduct experiments on three domain generalization datasets: **ColoredMNIST** (Arjovsky et al., 2020), **PACS** (Li et al., 2017) and **TerraIncognita** (Beery et al., 2018). We use the training-domain validation as defined in (Gulrajani & Lopez-Paz, 2020) for model selection.

As our method only modifies the mini-batch sampling strategy, we applied our proposed method along with four widely-used domain generalization baselines: **ERM** (Vapnik, 1998), **IRM** (Gulrajani & Lopez-Paz, 2020), **DRO** (Sagawa et al., 2019) and **CORAL** (Sun & Saenko, 2016), and compare the performance of using our method with using the usual random mini-batch sampling strategy.

Table 1. Out-of-domain test accuracy on ColoredMNIST, PACS and TerraIncognita dataset. Numbers are averaged over all test environments with 3 runs with standard deviation.

	Alg	CMNIST	PACS	TerraIncog
Random	ERM	51.5 ± 0.1	83.7 ± 0.1	46.1 ± 1.2
	IRM	42.5 ± 1.0	81.2 ± 0.4	39.3 ± 1.8
	DRO	51.9 ± 0.0	83.7 ± 0.4	43.6 ± 1.2
	CORAL	51.4 ± 0.1	84.8 ± 0.2	44.8 ± 0.3
Ours-Z	ERM	46.5 ± 0.9	**85.2** ± 0.3	47.1 ± 0.6
	IRM	44.6 ± 4.7	82.6 ± 0.3	40.1 ± 1.4
	DRO	49.2 ± 0.7	84.7 ± 0.6	41.6 ± 1.5
	CORAL	48.4 ± 1.7	84.3 ± 0.5	47.3 ± 0.4
Ours-$s^e(Z)$	ERM	**58.8** ± 0.5	**85.2** ± 0.4	**48.1** ± 0.3
	IRM	47.6 ± 1.6	82.1 ± 0.6	40.2 ± 2.9
	DRO	56.6 ± 2.2	84.3 ± 0.4	43.1 ± 0.6
	CORAL	58.6 ± 0.2	85.1 ± 0.4	46.2 ± 0.4

We use both $b(Z) = Z$ and $b(Z) = s^e(Z)$ to construct balanced mini-batches. For details, see Appendix C.

ColoredMNIST: In Table 1, $b(Z) = s^e(Z)$ significantly outperforms $b(Z) = Z$, and outperforms the random mini-batch sampling baseline by more than 5% (absolute) with all four classifiers. This is likely due to the relatively large noise (25%) in the label assignment: Z captures all other features except the ones directly determines the label, while $s^e(Z)$ only capture the features that has strong correlation with the label Y. i.e. the color information. As Z is finer and contains more information than $s^e(Z)$, it is likely that the closest match by $b(Z) = Z$ would be an image of the same digit with the same color. i.e. The matched example is likely to be an image labeled with the "wrong" class. This will hurt the performance as we up-sample the noise data.

PACS and TerraIncognita: Table 1 shows that our balanced mini-batch sampling method outperforms the random mini-batch sampling baseline on two real-world datasets. We observe that the performance difference between $b(Z) = s^e(Z)$ and $b(Z) = Z$ is not as large as that on the ColoredMNIST dataset. It is partially because these two datasets have less label noise than ColoredMNIST.

5. Related Work

A growing body of work has investigated the out-of-domain (OOD) generalization problem with causal modeling. One prominent idea is to learn invariant causal features that describe the true causal mechanism of interest across domains.

When multiple training domains are available, this can be approximated by enforcing some invariance conditions across the observed training domains by adding a regularization term to the usual empirical risk minimization (Arjovsky et al., 2020; Krueger et al., 2021; Bellot & van der Schaar, 2020; Wald et al., 2021). However, recent work claims that many of these approaches still fail to achieve the intended invariance property (Kamath et al., 2021; Rosenfeld et al., 2020; Guo et al., 2021), and thorough empirical study questions the true effectiveness of these domain generalization methods (Gulrajani & Lopez-Paz, 2020).

Some works propose to use an auxiliary variable different from the label instead of datasets from multiple domains to solve the OOD problem (Makar et al., 2022; Puli et al., 2022).Their methods are two-phased that first balance the train data distribution with the help of auxiliary variable and then add invariance regularizations on the training objective.

With appropriate assumptions of the train data distribution, some other OOD works propose to use variational autoencoder (VAE) to learn latent variables in the assumed causal graph (Liu et al., 2021; Lu et al., 2022), instead of using an observed auxiliary variable. The identifiability guarantee is usually based on the pioneer work on identifiable VAE by (Motiian et al., 2017).

Our method is based on the idea of both distribution balancing and latent variable learning. To better utilize the learned latent variable, we use a classic method for average treatment effect (ATE) estimation (Holland, 1986) – balancing score matching (Rosenbaum & Rubin, 1983). Recently, perfect match (Schwab et al., 2018) extends this method to individual treatment effect (ITE) estimation (Holland, 1986) by constructinng virtually randomized mini-batches with balancing score.

6. Conclusion

We propose a causality-based domain generalization method that samples balanced mini-batches to reduce spurious correlations. We show that our assumed data generation model with invariant causal mechanism can be identified up to sample transformations. We demonstrate theoretically that the balanced mini-batch is approximately sampled from a spurious free data distribution with the same causal mechanism under idea scenarios. Our experiments empirically show the effectiveness of our method on both semi-synthetic dataset and real-world datasets.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization, 2020.

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Bellot, A. and van der Schaar, M. Accounting for unobserved confounding in domain generalization. *arXiv preprint arXiv:2007.10653*, 2020.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of*

the American statistical Association, 112(518):859–877, 2017.

Dawid, A. P. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979. ISSN 00359246. URL http://www.jstor.org/stable/2984718.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

Guo, R., Zhang, P., Liu, H., and Kiciman, E. Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix. *arXiv preprint arXiv:2101.07732*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Holland, P. W. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. doi: 10.1080/01621459.1986.10478354. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kamath, P., Tangella, A., Sutherland, D. J., and Srebro, N. Does invariant risk minimization capture invariance? In *AISTATS*, 2021.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Liu, C., Sun, X., Wang, J., Tang, H., Li, T., Qin, T., Chen, W., and Liu, T.-Y. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34, 2021.

Lu, C., Wu, Y., Hernández-Lobato, J. M., and Schölkopf, B. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=-e4EXDWXnSn.

Mahajan, D., Tople, S., and Sharma, A. Domain generalization using causal matching. In *International Conference on Machine Learning*, pp. 7313–7324. PMLR, 2021.

Makar, M., Packer, B., Moldovan, D., Blalock, D., Halpern, Y., and D'Amour, A. Causally motivated shortcut removal using auxiliary labels. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 739–766. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/makar22a.html.

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5715–5725, 2017.

Pearl, J. *Causality*. Cambridge university press, 2009.

Puli, A. M., Zhang, L. H., Oermann, E. K., and Ranganath, R. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=12RoR2o32T.

Quiñonero-Candela, J., Sugiyama, M., Lawrence, N. D., and Schwaighofer, A. *Dataset shift in machine learning*. Mit Press, 2009.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41. URL https://doi.org/10.1093/biomet/70.1.41.

Rosenfeld, E., Ravikumar, P., and Risteski, A. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

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.

Schwab, P., Linhardt, L., and Karlen, W. Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks. *arXiv preprint arXiv:1810.00656*, 2018.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59, 2017. URL http://jmlr.org/papers/v18/16-011.html.

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks, 2014.

Vapnik, V. Statistical learning theory wiley. 1998.

Wald, Y., Feder, A., Greenfeld, D., and Shalit, U. On calibration and out-of-domain generalization. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=XWYJ25-yTRS.

A. Algorithm details

(a) As realized by ColoredMNIST (b) As realized by PACS

Figure 2. Annotated example causal graphs of two realizations of the joint distribution $p(X, Y, E)$.

Algorithm 1 Balanced Mini-batch sampling.

Input: $|\mathcal{E}_{\text{train}}|$ training datasets $\mathcal{D}^e = \{(x_i^e, y_i^e)\}_{i=1}^{N^e}$ sampled from distribution $P(X, Y|E = e)$ for all $e \in \mathcal{E}_{\text{train}}$, a balancing score $b^e(z_i)$ calculated for each training data point (x_i^e, y_i^e), and a distance metrics $d(\cdot, \cdot)$ that calculates the distance between two balancing scores

$D_{balanced} \leftarrow$ Empty

for $e \in \mathcal{E}_{\text{train}}$ **do**

  Randomly sample B data points D_{random}^e from $\mathcal{D}^e$

  Add D_{random}^e to $D_{balanced}$

  for $(x^e, y^e) \in D_{random}^e$ **do**

    Uniformly sample α different labels Y_{alt} from $\mathcal{Y} = \{1, 2, ..., m\}$ such that $y \neq y^e$ for all $y \in Y_{alt}$

    Suppose the balancing score of (x^e, y^e) is $b^e(z)$

    for $y \in Y_{alt}$ **do**

      Search across $\mathcal{D}^e$ for the data point (x_j^e, y_j^e) such that $y_j^e = y$ and has the smallest $d(b^e(z_j), b^e(z))$

      Add (x_j^e, y_j^e) to $D_{balanced}$

    end for

  end for

end for

B. Proofs

In this section, we give full proofs of the main theorems in the paper.

B.1. Balanced mini-batch sampling

B.1.1. Proof for Theorem 3.6

Our proof of all possible balancing scores is an extension of the proof of Theorem 2 from (Rosenbaum & Rubin, 1983), by generalizing the binary treatment to multiple treatments.

Proof. First, suppose the balancing score $b(Z)$ is finer than the propensity score $s(Z)$. By the definition of a balancing score (Definition 3.4) and Bayes' rule, we have:

$$p(Y|Z, b(Z)) = p(Y|b(Z)) \tag{11}$$

On the other hand, since $b(Z)$ is a function of Z, we have:

$$p(Y|Z, b(Z)) = p(Y|Z) \tag{12}$$

Equation (11) and Equation (12) give us $p(Y|b(Z)) = p(Y|Z)$. So to show $b(Z)$ is a balancing score, it is sufficient to show $p(Y|b(Z)) = p(Y|Z)$.

Let the y-th entry of $S(Z)$ be $s_y(Z) = p(Y = y|Z)$, then:

$$\mathbb{E}[s_y(Z)|b(Z)] = \int_{\mathcal{Z}} p(Y = y|Z = z)p(Z = z|b(Z))dz = p(Y = y|b(Z)) \tag{13}$$

But since $b(Z)$ is finer than $s(Z)$, $b(Z)$ is also finer than $s_y(Z)$, then

$$\mathbb{E}[s_y(Z)|b(Z)] = s_y(Z) \tag{14}$$

Then by Equation (13) and Equation (14) we have $P(Y = y|Z) = P(Y = y|b(Z))$ as required. So $b(Z)$ is a balancing score.

For the converse, suppose $b(Z)$ is a balancing score, but that $b(Z)$ is not finer than $s(Z)$. Then there exists z_1 and z_2 such that $s(z_1) \neq s(z_2)$, but $b(z_1) = b(z_2)$. By the definition of $s(\cdot)$, there exists y such that $P(Y = y|z_1) \neq P(Y = y|z_2)$. This means, Y and Z are not conditionally independent given $b(Z)$, thus $b(Z)$ is not a balancing score. Therefore, to be a balancing score, $b(Z)$ must be finer than $s(Z)$.

Note that $s(Z)$ is also a balancing score, since $s(Z)$ is also a function of itself.

□

B.1.2. Proof for Theorem 3.7

We provide a proof for Theorem 3.7, demonstrating the feasibility of balanced mini-batch sampling.

Proof. In Algorithm 1, by uniformly sampling a different labels such that $y \neq y^e$, we mean sample $Y_{\text{alt}} = \{y_1, y_2, ..., y_a\}$ by the following procedure:

$$\begin{aligned}
&y_1 \sim U\{1, 2, ..., m\} \setminus \{y_e\} \\
&y_2 \sim U\{1, 2, ..., m\} \setminus \{y_e, y_1\} \\
&\quad\vdots \\
&y_a \sim U\{1, 2, ..., m\} \setminus \{y_e, y_1, y_2 ... y_{a1}\}
\end{aligned}$$

Where U denotes the uniform distribution. Suppose $D_{\text{balanced}} \sim \hat{p}^B(X, Y)$, and data distribution $\mathcal{D}^e \sim p(X, Y|E = e), \forall e \in \mathcal{E}_{\text{train}}$.

Suppose we have an exact match every time we match a balancing score, then for all $e \in \mathcal{E}_{\text{train}}$, we have

$$\begin{aligned}
\hat{p}^B(Y|b^e(Z), E = e) &= \frac{1}{a+1}p(Y|b^e(Z), E = e) + \frac{1}{a+1}(1 - p(Y|b^e(Z), E = e)\frac{1}{m-1} + \\
&+ \frac{1}{a+1}(1 - p(Y|b^e(Z), E = e)(1 - \frac{1}{m-1})\frac{1}{m-2} + ... \\
&+ \frac{1}{a+1}(1 - p(Y|b^e(Z), E = e)(1 - \frac{1}{m-1})(1 - \frac{1}{m-2})...(1 - \frac{1}{m-a+1})\frac{1}{m-a} \\
&= \frac{1}{a+1}(\frac{a}{m-1} + \frac{m-a-1}{m-1}p(Y|b^e(Z), E = e))
\end{aligned}$$

By the definition of balancing score, $p(Y|Z, E = e) = p(Y|b^e(Z), E = e)$ and $\hat{p}^B(Y|Z, E = e) = \hat{p}^B(Y|b^e(Z), E = e)$, then we have

$$\hat{p}^B(Y|Z, E) = \frac{1}{a+1}(\frac{a}{m-1} + \frac{m-a-1}{m-1}p(Y|Z, E)) \tag{15}$$

When $a = m - 1$, we have $\hat{p}^B(Y|Z, E) = \frac{1}{m} = U\{1, 2, ..., m\}$, which means $\hat{p}^B(X, Y, Z) = p^B(X, Y, Z)$. i.e. D_{balanced} can be regarded as sampled from the balanced distribution p^B as defined in Definition 2.2.

□

B.1.3. Proof for Theorem 2.3

Here we give a proof of the minimax optimality of the Bayes optimal classifier trained on a balanced distribution.

Proof. The Bayes optimal classifier trained on a balanced distribution $p^B(X, Y)$ has $p_\psi(Y|X) = p^B(Y|X)$. Then consider the expected cross entropy loss of such classifier on an unseen test distribution p^e:

$$
\begin{aligned}
L^e(p^B(Y|X)) &= -\mathbb{E}_{p^e(X,Y)} \log p^B(Y|X) && (16)\\
&= -\mathbb{E}_{p^e(X,Y)} \log p^B(Y) + \mathbb{E}_{p^e(X,Y)} \log \frac{p^B(Y)}{p^B(Y|X)} && (17)\\
&= L^e(p^B(Y)) + \mathbb{E}_{p^e(X,Y,Z)} \left[\log \frac{p^B(Y)}{p^B(Y|X)}\right] && (18)\\
&= L^e(p^B(Y)) + \mathbb{E}_{p^e(Y,Z)} \left[\mathbb{E}_{p^B(X|Y,Z)} \left[\log \frac{p^B(Y)}{p^B(Y|X)}\right]\right] && (19)\\
&= L^e(p^B(Y)) + \mathbb{E}_{p^e(Y,Z)} \left[\mathbb{E}_{p^B(X|Y,Z)} \left[\log \frac{p^B(Y|Z)}{p^B(Y|X,Z)}\right]\right] && (20)\\
&= L^e(p^B(Y)) - \mathbb{E}_{p^e(Y,Z)} KL[p^B(Y|X,Z)||p^B(Y|Z)] && (21)
\end{aligned}
$$

- Equation (16) is the definition of cross entropy loss.
- Equation (18) is obtained by $Y \perp\!\!\!\perp_B Z$ and $Y \perp\!\!\!\perp_B Z|X$.

Thus we have the cross entropy loss of $p^B(X, Y)$ in any environment e is smaller than that of $p^B(Y) = \frac{1}{m}$ (random guess):

$$L^e(p^B(Y|X)) - L^e(p^B(Y)) \leq -\mathbb{E}_{p^e(Y,Z)} KL[p^B(Y|X,Z)||p^B(Y|Z)] \leq 0$$

Which means:

$$\max_{e' \in \mathcal{E}} \left[L^{e'}(p^B(Y|X)) - L^{e'}(p^B(Y))\right] \leq 0$$

That is, the performance of $p^B(X, Y)$ is at least as good as random guess in any environment. Since we make an assumption of the environment diversity, that is for any p^e with $Y \not\perp\!\!\!\perp_e Z$, there exist a environment e' such that $p^e(Y|X)$ performs worse than random guess. So we have:

$$\max_{e' \in \mathcal{E}} \left[L^{e'}(p^B(Y|X)) - L^{e'}(p^B(Y))\right] \leq 0 < \max_{e' \in \mathcal{E}} \left[L^{e'}(p^e(Y|X)) - L^{e'}(p^B(Y))\right]$$

Now we want to prove that $\forall e \in \mathcal{E},\ Y \perp\!\!\!\perp_e Z,\ Y \perp\!\!\!\perp_e Z|X,\ p^e(Y) = \frac{1}{m} \implies p^e(Y|X) = p^B(Y|X)$. For any $Z \in \mathcal{Z}$, we have:

$$p^e(Y|X) = p^e(Y|X,Z) = p^e(Y)\frac{p^e(X|Y,Z)}{\mathbb{E}_{p^e(Y|Z)}[p^e(X|Z,Y)]} = p^B(Y)\frac{p^B(X|Y,Z)}{\mathbb{E}_{p^B(Y)}[p^B(X|Z,Y)]} = p^B(Y|X,Z) = p^B(Y|X)$$

Thus we have the following minimax optimality:

$$p^B(Y|X) = \operatorname*{argmin}_{p_\psi \in \mathcal{F}} \max_{e \in \mathcal{E}} L^e(p_\psi(Y|X))$$

□

B.2. Latent Covariate Learning

B.2.1. Proof for Theorem 3.3

We now prove Theorem 3.3 setting up the identifiability of the necessary parameters that capture the spuriously correlated covariate features in the VAE. The proof of this Theorem is based on the proof of Theorem 1 in (Motiian et al., 2017), with the following modifications:

1. We use both E and Y as auxiliary variables.
2. We include Y in the causal mechanism of generating X by $X = \mathbf{f}(Y, Z) + \epsilon = \mathbf{f}_Y(Z) + \epsilon$.

Proof. **Step I.** In this step, we transform the equality of the marginal distributions over observed data into the equality of a noise-free distribution. Suppose we have two sets of parameters $\theta = (\mathbf{f}, \mathbf{T}, \lambda)$ and $\theta' = (\mathbf{f}', \mathbf{T}', \lambda')$ such that $p_\theta(X|Y, E = e) = p_{\theta'}(X|Y, E = e), \forall e \in \mathcal{E}_{\text{train}}$, then:

$$\int_{\mathcal{Z}} p_{\mathbf{T},\lambda}(Z|Y, E = e) p_{\mathbf{f}}(X|Z, Y) dZ = \int_{\mathcal{Z}} P_{\mathbf{T}',\lambda'}(Z|Y, E = e) p'_{\mathbf{f}}(X|Z, Y) dZ \tag{22}$$

$$\Rightarrow \int_{\mathcal{Z}} p_{\mathbf{T},\lambda}(Z|Y, E = e) p_\epsilon(X - \mathbf{f}_Y(Z)) dZ = \int_{\mathcal{Z}} p_{\mathbf{T}',\lambda'}(Z|Y, E = e) p_\epsilon(X - \mathbf{f}'_Y(Z)) dZ \tag{23}$$

$$\Rightarrow \int_{\mathcal{X}} p_{\mathbf{T},\lambda}(\mathbf{f}^{-1}(\bar{X})|Y, E = e) \text{vol} J_{\mathbf{f}^{-1}}(\bar{X}) p_\epsilon(X - \bar{X}) d\bar{X} = \int_{\mathcal{X}} p_{\mathbf{T}',\lambda'}(\mathbf{f}'^{-1}(\bar{X})|Y, E = e) \text{vol} J_{\mathbf{f}'^{-1}(\bar{X})} p_\epsilon(X - \bar{X}) d\bar{X} \tag{24}$$

$$\Rightarrow \int_{\mathbb{R}^d} \tilde{p}_{\mathbf{T},\lambda,\mathbf{f},Y,e}(\bar{X}) p_\epsilon(X - \bar{X}) d\bar{X} = \int_{\mathbb{R}^d} \tilde{p}_{\mathbf{T}',\lambda',\mathbf{f}',Y,e}(\bar{X} p_\epsilon(X - \bar{X}) d\bar{X}) \tag{25}$$

$$\Rightarrow (\tilde{p}_{\mathbf{T},\lambda,\mathbf{f},Y,e} * p_\epsilon)(X) = (\tilde{p}_{\mathbf{T}',\lambda',\mathbf{f}',Y,e} * P_{\mathcal{E}})(X) \tag{26}$$

$$\Rightarrow \mathscr{F}[\tilde{p}_{\mathbf{T},\lambda,\mathbf{f},Y,e}](\omega)\phi_\epsilon(\omega) = \mathscr{F}[\tilde{p}_{\mathbf{T}',\lambda',\mathbf{f}',Y,e}](\omega)\phi_\epsilon(\omega) \tag{27}$$

$$\Rightarrow \mathscr{F}[\tilde{p}_{\mathbf{T},\lambda,\mathbf{f},Y,e}](\omega) = \mathscr{F}[\tilde{p}_{\mathbf{T}',\lambda',\mathbf{f}',Y,e}](\omega) \tag{28}$$

$$\Rightarrow \tilde{p}_{\mathbf{T},\lambda,\mathbf{f},Y,e}(X) = \tilde{p}_{\mathbf{T}',\lambda',\mathbf{f}',Y,e}(X) \tag{29}$$

- In Equation (24), we denote the volume of a matrix $\mathbf{A}$ as $\text{vol}\mathbf{A} := \sqrt{\det \mathbf{A}^T \mathbf{A}}$. J denotes the Jacobian. We made the change of variable $\bar{X} = \mathbf{f}_Y(Z)$ on the left hand side and $\bar{X} = \bar{\mathbf{f}}_Y(Z)$ on the right hand side. Since $\mathbf{f}$ is injective, we have $\mathbf{f}^{-1}(\bar{X}) = (Y, Z)$. Here we abuse $\mathbf{f}^{-1}(\bar{X})$ to specifically denote the recovery of Z, i.e. $\mathbf{f}^{-1}(\bar{X}) = Z$.
- In Equation (25), we introduce

$$\tilde{p}_{\mathbf{T},\lambda,\mathbf{f},Y,e}(X) = p_{\mathbf{T},\lambda}(\mathbf{f}_Y^{-1}(X)|Y, E = e) \text{vol} J_{\mathbf{f}_Y^{-1}}(X) \mathbb{1}_{\mathcal{X}}(X) \tag{30}$$

  on the left hand side, and similarly on the right hand side.
- In Equation (26), we use $*$ for the convolution operator.
- In Equation (27), we use $\mathscr{F}[\cdot]$ to designate the Fourier transform, and the characteristic function of ϵ is $\phi_\epsilon = \mathscr{F}[p_\epsilon]$.
- In Equation (28), we dropped $\phi_\epsilon(\omega)$ from both sides as it is non-zero almost everywhere (by assumption (1) of the Theorem).

Step II. In this step, we remove all terms that are either a function of X or Y or e. By taking logarithm on both sides of Equation (29) and replacing $P_{\mathbf{T},\lambda}$ by its expression from Equation (3) we get:

$$\begin{aligned} &\log \text{vol} J_{\mathbf{f}^{-1}}(X) + \sum_{i=1}^{n} (\log Q_i(\mathbf{f}_i^{-1}(X)) - \log W_i^e(Y) + \sum_{j=1}^{k} \mathbf{T}_{i,j}(\mathbf{f}_i^{-1}(X)) \lambda_{i,j}^e(Y)) \\ =& \log \text{vol} J_{\mathbf{f}'^{-1}}(X) + \sum_{i=1}^{n} (\log Q'_i(\mathbf{f}_i'^{-1}(X)) - \log W_i'^e(Y) + \sum_{j=1}^{k} \mathbf{T}'_{i,j}(\mathbf{f}_i'^{-1}(X)) \lambda_{i,j}'^e(Y)) \end{aligned} \tag{31}$$

Let $(e_0, y_0), (e_1, y_1), ..., (e_{nk}, y_{nk})$ be the points provided by assumption (3) of the Theorem. We evaluate the above equations at these points to obtain $k+1$ equations, and subtract the first equation from the remaining k equations to obtain:

$$\begin{aligned}&\langle \mathbf{T}(\mathbf{f}^{-1}(X)), \lambda^{e_l}(y_l) - \lambda^{e_0}(y_0)\rangle + \sum_{i=1}^{n} \log \frac{W_i^{e_0}(y_0)}{W_i^{e_l}(y_l)} \\ =&\langle \mathbf{T}'(\mathbf{f}^{-1}(X)), \lambda'^{e_l}(y_l) - \lambda'^{e_0}(y_0)\rangle + \sum_{i=1}^{n} \log \frac{W_i'^{e_0}(y_0)}{W_i'^{e_l}(y_l)}\end{aligned} \tag{32}$$

Let $\mathbf{L}$ be the matrix defined in assumption (3) and $\mathbf{L}'$ similarly defined for λ' ($\mathbf{L}'$ is not necessarily invertible). Define $b_l = \sum_{i=1}^{n} \log \frac{W_i^{e_0'}(y_0) W_i^{e_l}(y_l)}{W_i^{e_0}(y_0) W_i^{e_L'}(y_l)}$ and $\mathbf{b} = [b_l]_{l=1}^{nk}$.

Then Equation (32) can be rewritten in the matrix form:

$$\mathbf{L}^T \mathbf{T}(\mathbf{f}^{-1}(X)) = \mathbf{L}'^T \mathbf{T}'(\mathbf{f}'^{-1}(X)) + \mathbf{b} \tag{33}$$

We multiply both sides of Equation (33) by $\mathbf{L}^{-T}$ to get:

$$\mathbf{T}(\mathbf{f}^{-1}(X)) = \mathbf{A}\mathbf{T}'(\mathbf{f}'^{-1}(X)) + \mathbf{c} \tag{34}$$

Where $\mathbf{A} = \mathbf{L}^{-T}\mathbf{L}'$ and $\mathbf{c} = \mathbf{L}^{-T}\mathbf{b}$.

Step III. To complete the proof, we need to show that $\mathbf{A}$ is invertible. By definition of $\mathbf{T}$ and according to Assumption (2), its Jacobian exists and is an $nk \times n$ matrix of rank n. This implies that the Jacobian of $\mathbf{T}' \circ \mathbf{f}'^{-1}$ exists and is of rank n and so is $\mathbf{A}$.

We distinguish two cases:

1. If $k = 1$, then $\mathbf{A}$ is invertible as $\mathbf{A} \in \mathbb{R}^{n \times n}$.

2. If $k > 1$, define $\bar{\mathbf{x}} = \mathbf{f}^{-1}(\mathbf{x})$ and $\mathbf{T}_i(\bar{x}_i) = (T_{i,1}(\bar{x}_i), ..., T_{i,k}(\bar{x}_i))$.

   Suppose for any choice of $\bar{x}_i^1, \bar{x}_i^2, ..., \bar{x}_i^k$, the family $(\frac{d\mathbf{T}_i(\bar{x}_i^1)}{d\bar{x}_i^1}, ..., \frac{d\mathbf{T}_i(\bar{x}_i^k)}{d\bar{x}_i^k})$ is never linearly independent. This means that $\mathbf{T}_i(\mathbb{R})$ is included in a subspace of $\mathbb{R}^k$ of dimension of most $k-1$. Let $\mathbf{h}$ be a non-zero vector that is orthogonal to $T_i(\mathbb{R})$. Then for all $x \in \mathbb{R}$, we have $\langle \frac{d\mathbf{T}_i(x)}{dx}, \mathbf{h}\rangle = 0$. By integrating we find that $\langle \mathbf{T}_i(x), \mathbf{h}\rangle = \text{const}$.

   Since this is true for all $x \in \mathbb{R}$ and for a $h \neq 0$, we conclude that the distribution is not strongly exponential. So by contradiction, we conclude that there exist k points $\bar{x}_i^1, \bar{x}_i^2, ...\bar{x}_i^k$ such that $(\frac{d\mathbf{T}_i(\bar{x}_i^1)}{d\bar{x}_i^1}, ..., \frac{d\mathbf{T}_i(\bar{x}_i^k)}{d\bar{x}_i^k})$ are linearly independent.

   Collect these points into k vectors $(\bar{\mathbf{x}}^1, ..., \bar{\mathbf{x}}^k)$ and concatenate the k Jacobians $J_{\mathbf{T}}(\bar{\mathbf{x}}^l)$ evaluated at each of those vectors horizontally into the matrix $\mathbf{Q} = (J_{\mathbf{T}}(\bar{\mathbf{x}}^1), ..., J_{\mathbf{T}}(\bar{\mathbf{x}}^k))$ and similarly define $\mathbf{Q}'$ as the concatentation of the Jacobians of $\mathbf{T}'(\mathbf{f}'^{-1} \circ \mathbf{f}(\bar{\mathbf{x}}))$ evaluated at those points. Then the matrix Q is invertible. By differentiating Equation (34) for each $\mathbf{x}^l$, we get:

$$\mathbf{Q} = \mathbf{A}\mathbf{Q}' \tag{35}$$

The invertibility of $\mathbf{Q}$ implies the invertibility of $\mathbf{A}$ and $\mathbf{Q}'$. This completes the proof.

□

C. Experiment Details

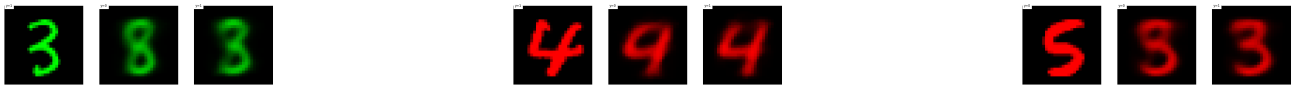

Figure 3. Reconstructed ColoredMNIST images from our VAE model. In each sub-figure, we infer Z from the leftmost image, then generate images with label 0 (middle) and 1 (right).

Baselines: Empirical risk minimization (ERM) is a default training scheme for most machine learning problems, merging all training data into one dataset and minimizing the training errors across all training domains. Invariant risk minimization (IRM) learns a data representation such that the optimal linear classifier on top of it is invariant across training domains. Group distributionally robust optimization (GroupDRO) performs ERM while increasing the weight of the environments with larger errors. Deep CORAL matches the mean and covariance of feature distributions across training domains.

We use a multi-layer perceptron based VAE (Kingma & Welling, 2013) to learn the latent covariate Z and we choose the conditional prior $p_\theta(Z|Y, E = e)$ to be a Gaussian distribution with diagonal covariance matrix. We also choose the noise distribution p_ϵ to be a Gaussian distribution with zero mean and fixed variance. For the architecture of the image classifiers (ERM (Vapnik, 1998), IRM (Gulrajani & Lopez-Paz, 2020), GroupDRO (Sagawa et al., 2019), CORAL (Sun & Saenko, 2016)), following the setting of DomainBed (Gulrajani & Lopez-Paz, 2020), we train a convolutional neural network from scratch for ColoredMNIST (Arjovsky et al., 2020) dataset, and use a pretrained ResNet50 (He et al., 2016) for PACS (Li et al., 2017) and TerraIncognita (Beery et al., 2018). Each experiment is repeated with 3 different random seeds.

We perform our experiments on the DomainBed codebase[2] and follow its default settings and hyperparameters. For all datasets, we set the number of matched examples to be $a = 1$.

Table 2. Out-of-domain test accuracy on ColoredMNIST dataset. Numbers are averaged over 3 runs with standard deviation.

Sampling	Alg	0.1	0.2	0.9	Avg
Random	ERM	71.8 ± 0.2	72.5 ± 0.1	10.1 ± 0.1	51.5 ± 0.1
	IRM	59.8 ± 1.5	58.1 ± 1.8	9.7 ± 0.0	42.5 ± 1.0
	DRO	72.7 ± 0.2	73.0 ± 0.2	10.0 ± 0.2	51.9 ± 0.0
	CORAL	71.4 ± 0.2	72.8 ± 0.1	9.9 ± 0.0	51.4 ± 0.1
Ours-Z	ERM	63.5 ± 2.2	66.0 ± 0.8	10.0 ± 0.1	46.5 ± 0.9
	IRM	54.4 ± 8.4	69.4 ± 7.5	9.9 ± 0.2	44.6 ± 4.7
	DRO	65.7 ± 1.5	67.1 ± 1.8	14.9 ± 4.0	49.2 ± 0.7
	CORAL	64.7 ± 0.8	65.5 ± 0.9	14.9 ± 4.1	48.4 ± 1.7
Ours-$s^e(Z)$	ERM	72.1 ± 0.1	71.2 ± 0.2	33.1 ± 1.3	58.8 ± 0.5
	IRM	69.8 ± 0.6	62.2 ± 5.6	10.9 ± 0.5	47.6 ± 1.6
	DRO	72.3 ± 0.2	71.4 ± 0.6	25.9 ± 6.7	56.6 ± 2.2
	CORAL	72.0 ± 0.6	71.8 ± 0.6	32.1 ± 0.7	58.6 ± 0.2

ColoredMNIST: The ColoredMNIST (Arjovsky et al., 2020) dataset is generated from the regular MNIST hand-written digits dataset (Lecun et al., 1998) by first assigning a binary label to indicate whether the digit is smaller than 5 to each MNIST image with some random noise, and then assigning a binary color (green or red) based on different correlations with the label across environments. Because of the injected noise in label assignment, the maximum accuracy a classifier can achieve is 75%. As shown in Table 2, there are in total 3 environments in the dataset, each of which has $p(\text{Color=Red}|Y = 1) = 0.1, 0.2, 0.9$, with 70,000 examples in each environment. In this case we have $|\mathcal{E}_{\text{train}}| = 2$ and $m = 2$. We also set $k = 1$ by fixing the variance of the conditional prior $p_\theta(Z|Y, E = e)$ to be 1. Then we take the maximum possible dimension of the latent Z, $n = 3$, according to the identifiability condition in Theorem 3.3. Note that as $m = 2$, the balanced mini-batches can be regarded as sampling from a balanced distribution with perfect match in balancing score and $a = 1$ as stated in Theorem 3.7.

Figure 3 shows three sets of reconstructed images with the same latent variable Z and different label Y using our VAE model. We can see that Z keeps the color feature and some style features, while the digit shape is changed with corresponding label Y.

[2]https://github.com/facebookresearch/DomainBed

As shown in Table 2, our proposed balanced mini-batch sampling method significantly outperforms random mini-batch sampling in the worst test environment (0.9) by 20% (absolute), while achieving comparable performance on the other two environments. Note that 0.9 is the hardest test environment as the train environments 0.1 and 0.2 are significantly different from it. While the accuracy on 0.9 is low, we are able to get around 75% (maximum) accuracy on the train domain validation sets.

Table 3. Out-of-domain test accuracy on each domain of the PACS dataset. Numbers are averaged over 3 runs. We use Gaussian distribution with $k = 2$ for latent covariate learning and $n = 64$.

Sampling	Alg	A	C	P	S	Avg
Random	ERM	85.0 ± 1.6	77.9 ± 2.2	95.4 ± 0.3	76.4 ± 1.1	83.7 ± 0.1
	IRM	81.4 ± 0.7	76.4 ± 1.3	96.6 ± 0.5	70.5 ± 2.0	81.2 ± 0.4
	DRO	83.6 ± 0.8	79.9 ± 1.2	96.4 ± 0.4	74.9 ± 0.3	83.7 ± 0.4
	CORAL	86.0 ± 0.6	77.6 ± 0.3	95.7 ± 0.1	80.1 ± 0.1	84.8 ± 0.2
Ours-Z	ERM	85.7 ± 0.7	80.2 ± 0.3	96.4 ± 0.2	78.5 ± 2.2	85.2 ± 0.3
	IRM	83.6 ± 1.2	74.7 ± 2.0	96.6 ± 0.3	75.6 ± 1.6	82.6 ± 0.3
	DRO	83.6 ± 1.2	80.1 ± 1.2	95.1 ± 0.5	80.2 ± 1.3	84.7 ± 0.6
	CORAL	82.2 ± 1.9	79.0 ± 1.1	96.2 ± 0.1	79.7 ± 0.8	84.3 ± 0.5
Ours-$s^e(Z)$	ERM	84.3 ± 0.5	79.8 ± 1.0	95.7 ± 0.4	80.9 ± 1.2	85.2 ± 0.4
	IRM	84.2 ± 1.7	73.8 ± 0.7	97.1 ± 0.4	73.4 ± 1.2	82.1 ± 0.6
	DRO	82.7 ± 0.9	79.0 ± 0.9	95.5 ± 0.3	80.0 ± 1.3	84.3 ± 0.4
	CORAL	84.0 ± 1.8	81.6 ± 0.9	94.2 ± 0.9	80.4 ± 1.1	85.1 ± 0.4

Table 4. Out-of-domain test accuracy on on each domain of the TerraIncognita dataset. Numbers are averaged over 3 runs. We use Gaussian distribution with $k = 2$ for latent covariate learning and $n = 29$.

Sampling	Alg	L100	L38	L43	L46	Avg
Random	ERM	50.6 ± 4.1	42.9 ± 2.6	55.8 ± 0.1	35.1 ± 2.0	46.1 ± 1.2
	IRM	34.0 ± 4.3	33.2 ± 9.1	50.4 ± 0.8	9.6 ± 1.1	39.3 ± 1.8
	DRO	50.0 ± 1.9	36.5 ± 4.8	56.3 ± 0.8	31.7 ± 3.1	43.6 ± 1.2
	CORAL	48.5 ± 2.9	37.9 ± 4.0	55.5 ± 2.1	37.3 ± 0.9	44.8 ± 0.3
Ours-Z	ERM	50.5 ± 2.9	45.1 ± 1.1	55.6 ± 0.9	37.4 ± 1.0	47.1 ± 0.6
	IRM	32.0 ± 3.9	40.8 ± 2.9	49.0 ± 1.5	38.7 ± 1.3	40.1 ± 1.4
	DRO	48.7 ± 3.1	23.3 ± 4.2	55.7 ± 1.0	38.5 ± 1.4	41.6 ± 1.5
	CORAL	54.9 ± 2.8	41.3 ± 0.9	55.2 ± 0.7	37.7 ± 1.7	47.3 ± 0.4
Ours-$s^e(Z)$	ERM	51.9 ± 0.4	45.8 ± 2.4	55.0 ± 1.0	39.8 ± 1.3	48.1 ± 0.3
	IRM	37.6 ± 7.2	43.0 ± 1.8	43.5 ± 3.8	36.9 ± 1.8	40.2 ± 2.9
	DRO	49.7 ± 2.7	34.6 ± 2.5	54.5 ± 1.2	33.4 ± 2.7	43.1 ± 0.6
	CORAL	50.1 ± 2.1	42.6 ± 2.1	54.7 ± 0.8	37.3 ± 2.5	46.2 ± 0.4

PACS and TerraIncognita: We also conduct experiments on two more realistic datasets, PACS (Li et al., 2017) and TerraIncognita (Beery et al., 2018). The PACS dataset has four environments, each indicates an image style: art, cartoons, photos and sketches, with 9,991 examples in total. There are 7 classes in the dataset, indicating the object in the image. The TerraIncognita dataset contains photographs of wild animals taken by camera traps at four different locations: L100, L38, L43 and L46, with 24,788 examples in total. There are 10 classes in the dataset, indicating the animal appears in the image.

In these two datasets, we do not require $|\mathcal{E}_{\text{train}}|$, m, n and k to explicitly satisfy the identifiability condition in Theorem 3.3, as the latent variable could require a larger dimensionality n to capture than the dataset allows. Also, as $m > 2$, $a = 1$ means we cannot completely eliminate the spurious correlation in the training dataset in the ideal scenario.

All experiments were conducted on NVidia A-100 and Titan X GPUs. The DomainBed codebase is released under an MIT license at https://github.com/facebookresearch/DomainBed. We utilize versions of the datasets ColoredMNIST, PACS, and TerraIncognita (Caltech Camera Traps) that are distributed within DomainBed, under the Creative Commons Attribution-NonCommercial 4.0, CopyLeft/No Rights Reserved, and Community Data License Agreement (CDLA) licenses respectively. Our code and corresponding instructions are included in the supplementary materials.