Representation via Representations: Domain Generalization via Adversarially Learned Invariant Representations

Zhun Deng *[1], Frances Ding [2], Cynthia Dwork[1], Rachel Hong[1], Giovanni Parmigiani [1], Prasad Patil [3], and Pragya Sur[1]

[1]Harvard University
[2]University of California, Berkeley
[3]Boston University

Abstract

We investigate the power of censoring techniques, first developed for learning *fair representations*, to address domain generalization. We examine *adversarial* censoring techniques for learning invariant representations from multiple "studies" (or domains), where each study is drawn according to a distribution on domains. The mapping is used at test time to classify instances from a new domain. In many contexts, such as medical forecasting, domain generalization from studies in populous areas (where data are plentiful), to geographically remote populations (for which no training data exist) provides fairness of a different flavor, not anticipated in previous work on algorithmic fairness.

We study an adversarial loss function for k domains and precisely characterize its limiting behavior as k grows, formalizing and proving the intuition, backed by experiments, that observing data from a larger number of domains helps. The limiting results are accompanied by non-asymptotic learning-theoretic bounds. Furthermore, we obtain sufficient conditions for good worst-case prediction performance of our algorithm on previously unseen domains. Finally, we decompose our mappings into two components and provide a complete characterization of invariance in terms of this decomposition. To our knowledge, our results provide the first formal guarantees of these kinds for adversarial invariant domain generalization.

1 Introduction

In gene expression analysis, as well as in much of high-throughput biology analyses on human populations, variation between studies can arise from the intrinsic biological heterogeneity of the populations being studied, or from technological differences in data acquisition. In turn both these types of variation can be shared across studies or not. For example, an algorithm for predicting whether a tumor will recur, trained on data obtained from the local population via a specific data-collection and processing method at a research hospital A, will typically not perform equally well on data collected at a research hospital B, using different data-collection techniques and serving a potentially different local population.

*Author names listed in alphabetical order. Corresponding authors: zhundeng@g.harvard.edu, dwork@seas.harvard.edu, pragya@seas.harvard.edu.

Of course, theoretically, B can train its own algorithm. This "siloization" is suboptimal for several reasons, from reduced statistical power, to wasteful allocation of research investment due to duplication of effort, or even reluctance to fund or publish such duplication. It is preferable to combine the data sources, potentially with some smart provision for domain variation. However, there will always be a new C to which the resulting algorithm will need to be applied – maybe the competition across town who did not collaborate at the development stage, or maybe a small, geographically isolated, population, far from any major medical research center. The ultimate goal, in the development of models with biomedical applications, is to provide accurate predictions for fully independent samples, originating from institutions and processed by laboratories that did not generate the training datasets. How can we transfer prediction capability to a new population?

This is a problem of *domain generalization*, the subject of intense study for nearly two decades [1, 4, 7, 9, 11–16, 19–22, 25, 26, 28]. Under the assumption that there is a common signal that provides a high quality predictor g^* for *all* populations, and given labeled training data from several populations, can this signal be learned even when it does not necessarily yield the best predictor for any given population? When does the presence of multiple training datasets improve the accuracy of this learning procedure?

Using tools developed for finding "fair" representations of individuals in which sensitive attributes such as sex or race have been censored [8, 17, 18, 29], we proceed from the following intuition: treating the domain as a sensitive attribute and training on multiple, highly diverse, populations, the learning algorithm is forced to disregard the idiosyncratic in favor of the universal, that is, to find a prediction rule based on a signal that is shared among all domains.

This work – domain generalization to unseen populations – provides a new dimension of fairness, *transferring the benefit of federal research dollars from preeminent bench to geographically remote bedside*, not anticipated in earlier work on learning fair representations.

Approach. We model the problem through the lens of a hierarchical Bayesian approach that is extensively used in applications. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the covariate space and $\mathcal{Y} = \{0, 1\}$ the outcome space. Let $\mathscr{D}$ denote a collection of probability distributions on $\mathcal{X} \times \mathcal{Y}$ and μ be a distribution supported on $\mathscr{D}$. The observed data arises through a hierarchical scheme—first, domains $\mathcal{D}_1, \cdots, \mathcal{D}_k$ are sampled i.i.d. from μ, and then random samples $S_i = \{x_{i,j}, y_{i,j}\}_{j=1}^{n_i}$ are drawn from each $\mathcal{D}_i$. We seek to train a classifier on the observed samples that performs well on any distribution from $\mathscr{D}$, even those from which no data have been observed. To this end, we adopt an adversarial censored learning approach. Simplifying slightly, for a mapping ϕ from the input covariate space to a representation space Z, a discriminator ψ_k that attempts to guess the source domain of $\phi^{-1}(z)$ for $z \in Z$, and a classifier f, we define an empirical adversarial loss function that increases with misclassifications by f and correct guesses by the discriminator (Equation 3). Our approach then tries to find the classifier f and encoding ϕ that minimizes this adversarial loss for the observed data. Our algorithm is adapted from [18], where it was used for the purposes of *fair representation learning*.

To study the performance of the proposed approach on a newly coming domain $\mathcal{D}_u \in \mathscr{D}$, it is crucial to understand the behavior of our adversarial loss in the limit of large k and n_i's. However, the structure of the discriminator changes with growing k. Thus, a crucial challenge lies in pinning down whether our loss admits a limit, and if so, what should be the limiting value? Additionally, even if we can characterize this limit, how would the proposed algorithm perform on an arbitrary $D_u \in \mathcal{D}$? This paper explores these key questions in detail.

Contributions. We obtain a precise characterization of the limit of our adversarial loss (Section 3.1). We address the challenges incurred by the dependence of the discriminator on k via a highly non-trivial geometric argument. We then provide non-asymptotic generalization error bounds for the empirical loss around its population counterpart; the form of the population version is naturally determined using the prior limiting result. We further establish consistency of loss function optimizers $\hat{f}_\lambda, \hat{\phi}_\lambda$, in the sense that, these converge (under an appropriate limit) to the corresponding optimizers of the population loss. Section 3.2 provides a characterization of the prediction performance of our algorithm on unseen domains that lie within bounded $\mathcal{H}$-divergence [3] of the seen ones. Section 3.3 decomposes our mappings ϕ into two components, and provides a complete characterization of *invariant* mappings (which defeat the discriminator) in terms of this decomposition. Extensive experimental results are summarized in Section 4.

Related Work. There are rich literatures of related work in computational learning theory. For lack of space we confine our discussion to a handful of works in domain generalization. In the earliest, kernel-based works on domain generalization [4,20], the learned classifier f receives at test time not just a single x drawn from a test distribution D_T, but (especially in [4]) a large, unlabeled sample from D_T together with a single additional test sample to be classified. To our knowledge, [20] is the first to assume a latent distribution on domains (as do we).

Three works are particularly aligned with our philosophical approach. [2] comes from a line of work, initiated in [24], on causal inference and predictive robustness, relying on a notion of probabilistic invariance. (See [5] for a survey.) [2] seeks data representations that elicit predictors satisfying certain invariance properties across the domains. This is framed as a penalized risk minimization problem, which is then solved using stochastic gradient descent. The theoretical guarantees rely on linearity assumptions [2, Theorem 8].

Inspired by [10], adversarial networks were introduced for fair representation learning in [8,18] and for domain generalization in [15,16]. [15] uses an autoencoder and introduces a Laplace prior on representations to encourage domain generalization. [16] employs an adversarial architecture very similar to ours, expanded with a subnetwork that seeks to minimize the discrepancy between $\mathbb{P}(X|Y)$ across the different domains, addressing differences in base rates among the training distributions. We provide theoretical insights not featured in [15, 16].

2 Formal setup

Recall our setting from Section 1. Throughout, we assume that $\mathscr{D}$ contains finitely many probability distributions, i.e. $\mathscr{D} = \{\mathcal{D}_1^*, \mathcal{D}_2^*, \cdots, \mathcal{D}_N^*\}$, and $\mathscr{D}^{\mathcal{X}}$ is the corresponding set of marginal distributions induced on $\mathcal{X}$. Define $\mathcal{D}_{1:k}$ to be the set of seen domains $\{\mathcal{D}_1, \ldots, \mathcal{D}_k\}$, and assign them distinct ID's $\{1, \ldots, k\}$. Let $S_{1:k}$ denote the collection of observed samples $\{S_1, \ldots, S_k\}$. Note that repeated sampling is possible here; for instance, we may have $\mathcal{D}_1 = \mathcal{D}_2 = \mathcal{D}_1^*$. Define $S_i^{\mathcal{X}} := \{x_{i,j}\}_{j=1}^{n_i}$ and $g(S_i^{\mathcal{X}}) := \{g(x_{i,j})\}_{j=1}^{n_i}$, for any function g on $\mathcal{X}$. For any function g and distribution $\mathcal{D}$, we use $g(\mathcal{D})$ to denote the distribution of $g(z)$, where $z \sim \mathcal{D}$. For any distribution $\mathcal{D}$ that admits a density function $p_{\mathcal{D}}$, let $\text{Supp}_{\mathcal{D}} := \{x | p_{\mathcal{D}}(x) > 0\}$. Finally, for any function $f : \mathbb{R}^m \to \mathbb{R}^n$, we represent $f(\cdot)$ using the n-dimensional vector $(f^{(1)}(\cdot), \ldots, f^{(n)}(\cdot))^\top$.

Algorithm. The samples $S_{1:k}$ are first passed through an encoder that produces a representation $\{\phi(S_i^{\mathcal{X}})\}_{i=1}^k$ of the input covariates. Here, ϕ is a representation mapping that belongs to some function class $\Phi = \{g | g : \mathbb{R}^d \to \mathbb{R}^s\}$. The output from the encoder is subsequently passed

through a discriminator ψ_k of the form

$$\psi_k(\cdot) = W\zeta(\cdot) + B, \tag{1}$$

where $\zeta : \mathbb{R}^s \to \mathbb{R}^p$ lies in some function class Υ, $W \in \mathbb{R}^{k\times p}$ and $B \in \mathbb{R}^k$. We further denote $W = (w_1, w_2, \cdots, w_k)^\top, B = (b_1, \cdots, b_k)^\top$, where $w_i \in \mathbb{R}^p, b_i \in \mathbb{R}$. Thus, the discriminator comprises a base structure ζ followed by a linear transformation, and effectively maps each input in $\mathbb{R}^s$ to k unnormalized weights. For each input $\phi(x_{i,j})$, the ℓ-th entry in the normalized version of the output $\psi_k(\phi(x_{i,j})) \in \mathbb{R}^k$ should be viewed as the discriminator's estimate of the probability that the pre-image $\phi^{-1}(\phi(x_{i,j}))$ was drawn from the seen domain with ID ℓ. Finally, define $\pi_k(\cdot)$ to be the operation that maps an input vector w to the index of the entry with maximal weight. If multiple entries achieve the maximal weight, π_k chooses uniformly among the corresponding indices. Simultaneously, a predictor is trained on the encoded representations $\{\phi(S_i^{\mathcal{X}})\}_{i=1}^k$ and produces labels in the outcome space. Denote the predictor class by $\mathcal{F} = \{f \mid f : \mathbb{R}^s \mapsto \mathcal{Y}\}$.

Loss function. The encoder, discriminator and predictor will be simultaneously trained using a loss function that comprises two components: (a) the loss corresponding to the predictor $L_{\text{pred}}(\mathcal{D}_{1:k}, f, \phi) = (1/k)\sum_{i=1}^k \mathbb{P}_{(x,y)\sim\mathcal{D}_i}(f(\phi(x)) \neq y)$, (b) the loss corresponding to the discriminator or adversary $L_{\text{adv}}(\mathcal{D}_{1:k}, \phi, \psi_k) = \sum_{i=1}^k \mathbb{P}_{x\sim\mathcal{D}_i^{\mathcal{X}}}(\pi_k \circ \psi_k(\phi(x)) = i)$. The form of these loss functions is inspired from [18]. Define

$$L(\mathcal{D}_{1:k}, f, \phi, \psi_k; \lambda) = L_{\text{pred}}(\mathcal{D}_{1:k}, f, \phi) + \lambda L_{\text{adv}}(\mathcal{D}_{1:k}, \phi, \psi_k), \tag{2}$$

where $\lambda > 0$ is a tuning parameter, and the corresponding empirical version

$$L(S_{1:k}, f, \phi, \psi_k; \lambda) = \frac{1}{k}\sum_{i=1}^k \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbb{1}\{f(\phi(x_{i,j})) \neq y_{i,j}\} + \lambda \sum_{i=1}^k \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbb{1}\{\pi_k \circ \psi_k(\phi(x_{i,j})) = i\}, \tag{3}$$

where $\mathbb{1}$ is the indicator function. We seek to optimize the aforementioned loss to obtain

$$(\hat{f}_\lambda, \hat{\phi}_\lambda) = \arg \inf_{f\in\mathcal{F}, \phi\in\Phi} \sup_{\psi_k\in\Psi_k} L(S_{1:k}, f, \phi, \psi_k; \lambda). \tag{4}$$

The infimum aims to maximize accuracy of the predictor, whereas the supremum ensures the performance of the discriminator is minimized. The final predictor for any test datapoint $x \sim \mathcal{D}$ where $\mathcal{D} \sim \mu$, is then given by $\hat{y} := \hat{f}_\lambda(\hat{\phi}_\lambda(x))$.

Remark 2.1. *Recall the definition of $\mathcal{H}$-Divergence [3]: let $\mathcal{H}$ be a class of binary classifiers, then $\mathcal{H}$-divergence between distributions $\mathcal{D}$ and $\mathcal{D}'$ over $\mathbb{R}^d$ is defined as*

$$D_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = \sup_{h\in\mathcal{H}} |\mathbb{P}_{x\sim\mathcal{D}}(h(x) = 1) - \mathbb{P}_{x\sim\mathcal{D}'}(h(x) = 1)|.$$

In the case of $k = 2$, if we choose $\mathcal{H} = \{\pi_2 \circ \psi_2(\phi(\cdot)) : \psi_2 \in \Psi_2, \phi \in \Phi\}$, then we can see that

$$\sum_{i=1}^2 \mathbb{P}_{x\sim\mathcal{D}_i^{\mathcal{X}}}(\pi_2 \circ \psi_2(\phi(x)) = i) = 1 + \mathbb{P}_{x\sim\mathcal{D}_1^{\mathcal{X}}}(\pi_2 \circ \psi_2(\phi(x)) = 1) - \mathbb{P}_{x\sim\mathcal{D}_2^{\mathcal{X}}}(\pi_2 \circ \psi_2(\phi(x)) = 1).$$

As a result,

$$\sup_{\psi_2\in\Psi_2} \sum_{i=1}^2 \mathbb{P}_{x\sim\mathcal{D}_i^{\mathcal{X}}}(\pi_2 \circ \psi_2(\phi(x)) = i) = 1 + d_{\mathcal{H}}(\phi(\mathcal{D}_1^{\mathcal{X}}), \phi(\mathcal{D}_2^{\mathcal{X}})).$$

Our loss is a natural generalization of $\mathcal{H}$- divergence and has a straightforward interpretation – how best can the discriminator distinguish the images from the k domains. We elucidate this further in Section 3.1.

For simplicity, throughout the paper, we only consider function classes for which the infimum and supremum can be achieved, and therefore replace inf, sup in (4) by min, max respectively. In our experiments, the function classes $\mathcal{F}, \Phi, \Psi_k$ are taken to be neural networks with specific architectures.

3 Theoretical Results

3.1 Learning theoretic analysis

To obtain a learning theoretic analysis of $L(S_{1:k}, f, \phi, \psi_k; \lambda)$, it is crucial to understand its limiting behavior when the sample sizes n_i and number of seen domains k diverge. It is clear that when every $n_i \to \infty$, $L(S_{1:k}, f, \phi, \psi_k; \lambda) \to L(\mathcal{D}_{1:k}, f, \phi, \psi_k; \lambda)$. Furthermore, $L_{\text{pred}}(\mathcal{D}_{1:k}, f, \phi) \to \mathbb{E}_{\mathcal{D}\sim\mu}[\mathbb{P}_{(x,y)\sim\mathcal{D}}(f(\phi(x)) \neq y)]$ as $k \to \infty$. Thus, we focus on studying the limiting behavior of $\max_{\psi_k\in\Psi_k} L_{\text{adv}}(\mathcal{D}_{1:k}, \phi, \psi_k)$ when k diverges. Denote the density function of $\zeta(\phi(\mathcal{D}_i^{*\mathcal{X}}))$ by $\rho_i^{\zeta,\phi}(\cdot)$.

Assumption 3.1 (Continuity)**.** *Every $\zeta \in \Upsilon$ and $\phi \in \Phi$ is continuous almost everywhere (a.e.). Besides, for all $\mathcal{D}_i^{*\mathcal{X}} \in \mathscr{D}$,$\zeta \in \Upsilon$ and $\phi \in \Phi$, $\rho_i^{\zeta,\phi}(\cdot)$ is continuous a.e. and $\text{Supp}_{\zeta(\phi(\mathcal{D}_i^{*\mathcal{X}}))}$ has non-zero volume in $\mathbb{R}^p$.*

Notice fully connected neural networks with ReLU activation functions are continuous a.e., since the ReLU activation function is discontinuous only at 0.

Theorem 3.1. *Suppose μ is a probability distribution supported on a finite set of distributions $\mathcal{D} := \{\mathcal{D}_1^*, \mathcal{D}_2^*, \cdots, \mathcal{D}_N^*\}$, each of which is a distribution over $\mathcal{X}\times\mathcal{Y}$. Further, let $\mathcal{D}_1, \cdots, \mathcal{D}_k \overset{\text{iid}}{\sim} \mu$ with corresponding study IDs $\{1, \ldots, k\}$. Then under Assumption 3.1, for any $\phi \in \Phi$,*

$$\lim_{k\to\infty} \max_{\psi_k\in\Psi_k} \sum_{i=1}^{k} \mathbb{P}_{x\sim\mathcal{D}_i^{\mathcal{X}}}(\pi_k \circ \psi_k(\phi(x)) = i) = \sup_{\cup_i A_i=\mathbb{R}^p, A_i\cap A_j=\emptyset, \zeta\in\Upsilon} \sum_{i=1}^{N} \mathbb{P}_{x\sim\mathcal{D}_i^{*\mathcal{X}}}(\zeta(\phi(x)) \in A_i). \quad (5)$$

The probabilities on the LHS (left hand side) are taken w.r.t. marginal covariate distributions of the seen domains $\mathcal{D}_i^{\mathcal{X}}$, which may contain repeats from $\mathscr{D}$. But the RHS contains probabilities w.r.t. the corresponding marginals of all distributions in $\mathscr{D}$. Speaking intuitively, Theorem 3.1 says that, for *every* $\phi \in \Phi$, as k grows so that (1) the encodings $\phi(\mathcal{D}_i^{\mathcal{X}})$ contain repeated instances of every element from $\mathscr{D}$, and (2) the structure of the last layer of the discriminator changes with k, the chance that the adversary accurately guesses the IDs of the encoded inputs is the same as the chance that the encoding $\phi(\cdot)$ itself maps the true distributions $\mathcal{D}_i^{*\mathcal{X}}$ to N disjoint parts of the space.

Theorem 3.1 provides further insights into our loss function (2) and the behavior of our algorithm. Since the result holds for any $\phi \in \Phi$, when k is large our algorithm effectively finds

$$(\hat{f}_\lambda, \hat{\phi}_\lambda) \approx \arg\min_{f,\phi}\{L_{\text{pred}}(\mathcal{D}_{1:k}, f, \phi) + \lambda \sup_{\cup_i A_i=\mathbb{R}^p, A_i\cap A_j=\emptyset, \zeta\in\Upsilon} \sum_{i=1}^{N} \mathbb{P}_{x\sim\mathcal{D}_i^{*\mathcal{X}}}(\zeta(\phi(x)) \in A_i)\}.$$

The second term is minimized for an encoding $\phi(\cdot)$ that maps the distributions $\mathcal{D}_i^{\star\mathcal{X}}$ to similar images, so that the adversary finds it difficult to guess the true IDs of the input covariates. (The prediction part of the loss discourages the trivial mapping $\forall x,\ \phi(x) = z$ for some arbitrary z.)

The limit in (5) may be viewed as a measure of dissimilarity of the *set* $\{\phi(\mathcal{D}_i^{\star\mathcal{X}})\}_{i=1}^N$. In fact, consider a setting where the supremum over $\zeta \in \Upsilon$ in the RHS of (5) is achieved and that the maximizer,

$$\zeta^\star = \text{argmax}_{\zeta\in\Upsilon} \sum_{i=1}^N \mathbb{P}_{x\sim\mathcal{D}_i^{*\mathcal{X}}}(\zeta(\phi(x)) \in A_i)$$

is unique. Then, it is not hard to see that

$$\sup_{\cup_i A_i=\mathbb{R}^p, A_i\cap A_j=\emptyset} \sum_{i=1}^N \mathbb{P}_{x\sim\mathcal{D}_i^{*\mathcal{X}}}(\zeta^\star(\phi(x)) \in A_i) \geqslant 1,$$

where equality holds iff $\zeta^\star(\phi(\mathcal{D}_1^{*\mathcal{X}})) = \cdots = \zeta^\star(\phi(\mathcal{D}_N^{*\mathcal{X}}))$.

If Υ contains the identity mapping, then

$$\sup_{\cup_i A_i=\mathbb{R}^p, A_i\cap A_j=\emptyset} \sum_{i=1}^N \mathbb{P}_{x\sim\mathcal{D}_i^{*\mathcal{X}}}(\zeta^\star(\phi(x)) \in A_i) = 1$$

iff $\phi(\mathcal{D}_1^{*\mathcal{X}}) = \cdots = \phi(\mathcal{D}_N^{*\mathcal{X}})$. That is, the limit in (5) is minimized iff $\{\phi(\mathcal{D}_i^{\star\mathcal{X}})\}_{i=1}^N$ are identical.

To the best of our knowledge, such a precise understanding of the adversarial loss, as illuminated by Theorem 3.1, has so far eluded prior literature, and may be of independent interest for invariant representation learning [27]. The fact that the structure of ψ_k changes with k presents significant challenges for the proof. We address this issue with a highly non-trivial geometric argument (Section A, Supplementary). Speaking informally, we cover the space $\mathbb{R}^p$ by grid cells such that, as k increases, the number of cells grows and each cell becomes increasingly refined. As the cells grow finer, each one can be associated with an element from $\mathscr{D}$ according to the distribution among $\{\zeta(\phi(\mathcal{D}_i^{\star\mathcal{X}}))\}_{i=1}^N$ whose density in the cell is largest (ignoring ties). Then the final layer can be chosen such that, for every cell, the adversary assigns the highest weight to the corresponding distribution.

Remark 3.1. *For $N = 2$, the limit can be related to the total variation distance since*

$$\sup_{\cup_i A_i=\mathbb{R}^p, A_i\cap A_j=\emptyset} \sum_{i=1}^2 \mathbb{P}_{x\sim\mathcal{D}_i^{*\mathcal{X}}}(\zeta^*(\phi(x)) \in A_i) = TV(\zeta^*(\phi(\mathcal{D}_1^{*\mathcal{X}})), \zeta^*(\phi(\mathcal{D}_2^{*\mathcal{X}}))) + 1.$$

Non-asymptotic generalization bounds. We now turn to bounding the generalization error of $L(S_{1:k}, f, \phi, \psi_k; \lambda)$. To this end, a few key quantities are introduced next. Define

$$\begin{aligned} L(\mathscr{D}, f, \phi; \lambda) &= \mathbb{E}_{\mathcal{D}\sim\mu}[\mathbb{P}_{(x,y)\sim\mathcal{D}}(f(\phi(x)) \neq y)] \\ &\quad + \sup_{\cup_i A_i=\mathbb{R}^p, A_i\cap A_j=\emptyset, \zeta\in\Upsilon} \lambda \sum_{i=1}^N \mathbb{P}_{x\sim\mathcal{D}_i^{*\mathcal{X}}}(\zeta(\phi(x)) \in A_i). \end{aligned} \tag{6}$$

Definition 3.1 (Grid Cells)**.** *For $n \in \mathbb{N}^+, B \in \mathbb{R}^+$, define $G(n, B)$ to be the set*

$$G(n, B) = \{I_{i_1} \times I_{i_2} \times \cdots \times I_{i_p}, i_j \in \{1, \cdots, n\}\ \forall\ j\},$$

where $I_{i_j} = [-B + 2(i_j - 1)B/n, -B + 2i_jB/n]$.

The elements in $G(n,B)$ form a partition of $[-B,B]^p$ and the intersection of every pair of elements has volume 0 in $\mathbb{R}^p$. Now, let H_k be a collection of distributions in $\mathscr{D}$ that receive high μ-probability in the following sense, $H_k := \{\mathcal{D}_i \in \mathscr{D} : \mu(\mathcal{D}_i) \geqslant 1/k^{1/4}\}$. Define $T_i^{\zeta,\phi}$ to be the set of points in $\mathbb{R}^p$ where the density $\rho_i^{\zeta,\phi}$ is maximized (up to ties), that is,

$$T_i^{\zeta,\phi} := \{z \in \mathbb{R}^p : \rho_i^{\zeta,\phi}(z) > \rho_j^{\zeta,\phi}(z), \text{ for all } j < i \text{ and } \rho_i^{\zeta,\phi}(z) \geqslant \rho_j^{\zeta,\phi}(z), \text{ for } j \geqslant i\}.$$

It is easy to see that $\{T_i^{\zeta,\phi}\}_{i=1}^N$ form a partition of $\mathbb{R}^p$. Furthermore, let $M_{i,1}^{\zeta,\phi}(n,B)$ denote the collection of grid cells on the boundary of $T_i^{\zeta,\phi}$, and $M_{i,2}^{\zeta,\phi}(n,B)$ denote those in the interior.

$$M_{i,1}^{\zeta,\phi}(n,B) = \{g \in G(n,B) | g \nsubseteq T_i^{\zeta,\phi}, g \cap T_i^{\zeta,\phi} \neq \emptyset\}$$
$$M_{i,2}^{\zeta,\phi}(n,B) = \{g \in G(n,B) | g \subseteq T_i^{\zeta,\phi}\}.$$

Assumption 3.2 (Boundedness)**.** *There exists a constant B_ρ and function $B(\cdot)$ s.t. for any $\varepsilon > 0$, $\sup_{\zeta,\phi} \sum_{i=1}^N \mathbb{P}_{x\sim\mathcal{D}_i^*} x(\|\zeta(\phi(x))\|_2 \geqslant B(\varepsilon)) \leqslant \varepsilon$ and $\sup_{z,\zeta,\phi,i} |\rho_i^{\zeta,\phi}(z)| \leqslant B_\rho$.*

Assumption 3.3 (Bounded VC-dimensions)**.** *Assume that the function classes $\Lambda = \{f \circ \phi | f \in \mathcal{F}, \phi \in \Phi\}$ and $\Xi = \{\mathbb{1}\{w_1^\top \zeta(\phi(x)) + b_1 > w_2^\top \zeta(\phi(x)) + b_2\} | w_i \in \mathbb{R}^p, b_i \in \mathbb{R}, \zeta \in \Upsilon, \phi \in \Phi, i = 1,2\}$ have VC-dimensions $\mathcal{V}_\Lambda$ and $\mathcal{V}_\Xi$ respectively.*

Note that in Assumption 3.3, the VC dimension condition on Ξ is on two nodes instead of k.

Theorem 3.2. *Consider the setting of Theorem 3.1, and define $m_k := \lceil k^{\frac{3}{4}} - \sqrt{(k\log(|H_k|) + k^{\frac{3}{4}})/\sqrt{2}} \rceil$. Under Assumptions 3.1-3.3, there exists a universal constant c, s.t. for any $t_1, t_2 > 0$, w.p. at least $1 - e^{-k^{1/4}} - \sum_{i=0}^{k-1} 4e^{-n_i t_1^2} - 2Ne^{-2kt_2^2}$,*

$$\max_{f\in\mathcal{F},\phi\in\Phi} |\max_{\psi_k\in\Psi_k} L(S_{1:k}, f, \phi, \psi_k; \lambda) - L(\mathscr{D}, f, \phi; \lambda)| \leqslant (1+k\lambda)t_1 + \frac{2\lambda}{\sqrt{k}} + N \cdot t_2 + I + II + III, \tag{7}$$

where $\mathcal{V}_{\mathcal{C}(k)} = k\mathcal{V}_\Xi(\log(\mathcal{V}_\Xi))^2$, and

$$I = \lambda \max\{N - |H_k|, 0\}, \quad II = \frac{2\lambda B_\rho \left(B(\frac{1}{\sqrt{k}})\right)^p}{\lfloor m_k^{1/p} \rfloor^p} \sum_{i\in H_k} \sup_{\zeta,\phi} |M_{i,1}^{\zeta,\phi}(\lfloor m_k^{1/p} \rfloor, B(\frac{1}{\sqrt{k}}))|,$$
$$III = \frac{2c}{k} \sum_{i=1}^k \frac{k\sqrt{\mathcal{V}_{\mathcal{C}(k)} \log(\frac{n_i}{\mathcal{V}_{\mathcal{C}(k)}})} + \sqrt{\mathcal{V}_\Lambda \log(\frac{n_i}{\mathcal{V}_\Lambda})}}{\sqrt{n_i}}.$$

We now proceed to analyze the bound in (7). Note that III vanishes when $\min_i n_i = \Omega(k^\alpha)$ for $\alpha \geqslant 2$, whereas I is small when k is much larger than N.

For II, note that for k much larger than N, $\log(|H_k|)$ is negligible compared to $\sqrt{k}$, so that $m_k = \Omega(k^{3/4})$. Now for fixed B, when k is large, $G(\lfloor m_k^{1/p} \rfloor, B)$ shrinks in volume. In settings where the union of the grid cells in $M_{i,2}^{\zeta,\phi}(\lfloor m_k^{1/p} \rfloor, B)$ approximates $T_i^{\zeta,\phi}$ well enough with growing k, $M_{i,1}^{\zeta,\phi}(\lfloor m_k^{1/p} \rfloor, B)$ contains negligible number of grid cells compared to $M_{i,2}^{\zeta,\phi}(\lfloor m_k^{1/p} \rfloor, B)$, leading to

$$\sum_{i\in H_k} \sup_{\zeta,\phi} |M_{i,1}^{\zeta,\phi}(\lfloor m_k^{1/p} \rfloor, B)| = o(\lfloor m_k^{1/p} \rfloor^p).$$

We defer the readers to the Supplementary Section A for specific examples demonstrating this phenomenon. This continues to hold when B is replaced by $B(1/\sqrt{k})$ if the latter grows slowly with increasing k. Since $B(\cdot)$ is related to the tails of the distributions $\zeta(\phi(\mathcal{D}_i^{*\mathcal{X}}))$, we are able to control this term in specific examples. For instance, if all distributions in $\mathscr{D}$ are sub-Gaussian with sub-gaussian norm bounded by some constant $\sigma_{\max}$, then $B(1/\sqrt{k}) = O(\sqrt{\log k})$. Together, this means that II is also small when k is sufficiently large. Thus, Theorem 3.2 demonstrates that observing samples from *more domains helps in generalization.*

Consistency. Theorem 3.2 provides conditions on k and n_i, $i \in [k]$, under which the empirical loss function, when evaluated at the estimates $(\hat{f}_\lambda, \hat{\phi}_\lambda)$, will be close to its population counterpart w.h.p. Here we seek to establish that these estimates, in fact, well approximate the minimizers of the population loss. Since we impose no assumptions on the specific distributional forms of the seen domains, this is hard to prove in such generality. We will therefore establish this under a curvature condition on the population loss that is slightly weaker than strong convexity.

Assumption 3.4 (Well-separation)**.** *Denote $\mathcal{M}^*_{\mathcal{F},\Phi} \subseteq \mathcal{F} \times \Phi$ to be the set of minimizers of $L(\mathscr{D}, f, \phi; \lambda)$. For a metric $dist(\cdot,\cdot)$ on the function class $\mathcal{F} \times \Phi$, there exists a function $U(\cdot;\lambda) : \mathbb{R} \to \mathbb{R}^+$ satisfying $\lim_{\varepsilon \to 0} U(\varepsilon;\lambda) \to 0$, such that for any $\varepsilon > 0$*

$$\inf_{\xi \in \mathcal{F}\times\Phi:\ \inf_{z \in \mathcal{M}^*_{\mathcal{F},\Phi}} dist(\xi,z) \geqslant U(\varepsilon;\lambda)} |L(\mathscr{D}, \xi; \lambda) - \min_{f \in \mathcal{F}, \phi \in \Phi} L(\mathscr{D}, f, \phi; \lambda)| \geqslant \varepsilon.$$

Theorem 3.3. *Under Assumption 3.1-3.4, almost surely,*

$$\inf_{z \in \mathcal{M}^*_{\mathcal{F},\Phi}} dist((\hat{f}_\lambda, \hat{\phi}_\lambda), z) \leqslant U(2\Gamma; \lambda),$$

where Γ equals the RHS of (7).

3.2 Generalization to unseen domains

Theorem 3.3 establishes that, under the aforementioned conditions, our proposed classifier $\hat{f}_\lambda(\hat{\phi}_\lambda(\cdot))$ minimizes the population loss $L(\mathscr{D}, f, \phi; \lambda)$. However, this loss is a penalized version of the expected prediction error under μ. Naturally, the results from the preceding section fail to capture the behavior of our classifier on an arbitrary domain from $\mathscr{D}$. We now address this problem, showing that such a worst case characterization is possible if elements in $\mathscr{D}$ are well-represented under μ—that is, every domain in $\mathscr{D}$ is close to at least one domain that receives relatively high μ-probability.

Assumption 3.5 (Well-represented)**.** *There exists constants $0 < p_l < 1$ and $\delta > 0$, s.t. for any $\mathcal{D} \in \mathscr{D}$ with $\mu(\mathcal{D}) > 0$, $\exists \mathcal{D}' \in \mathscr{D}$ with $\mu(\mathcal{D}') \geqslant p_l$ and $d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') \leqslant \delta$, where $\mathcal{H} = \mathcal{F} \times \Phi$.*

Theorem 3.4. *Under Assumptions 3.1, 3.3 and 3.5, w.p. at least $1 - \exp(-k^2 p_l^2/2)/p_l - \sum_{i=1}^k 4e^{-n_i t^2}$ over the randomness in $S_{1:k}$ and $\mathcal{D}_{1:k}$, for any $\mathcal{D}_u \in \mathscr{D}$ and all $f \in \mathcal{F}, \phi \in \Phi$,*

$$\mathbb{P}_{(x,y)\sim\mathcal{D}_u}(f(\phi(x)) \neq y) \leqslant \frac{2}{p_l}\Big(\hat{\beta}(f,\phi) + t + c\sqrt{\frac{\mathcal{V}_\Lambda \log(n_i/\mathcal{V}_\Lambda)}{n_i}}\Big) + \delta,$$

where $\hat{\beta}(f,\phi) = \sum_{i=1}^k \sum_{j=1}^{n_i} \mathbb{I}\{f(\phi(x_{i,j})) \neq y_{i,j}\}/(kn_i)$. Moreover, this holds even if $|\mathscr{D}|$ is countably infinite.

3.3 Characterization of invariant representation mappings

Definition 3.2. *An element $\phi \in \Phi$ is said to be an invariant representation mapping for a collection of k domains $\tilde{\mathcal{D}}_1, \ldots, \tilde{\mathcal{D}}_k \in \mathscr{D}$ and for some $\epsilon > 0$, if* $\sup_{\psi_k \in \Psi_k} \sum_{i=1}^{k} \mathbb{P}_{x \sim \tilde{\mathcal{D}}_i^{\mathcal{X}}}(\pi_k \circ \psi_k(\phi(x)) = i) \leqslant \varepsilon$.

Recall that the range of any $\phi \in \Phi$ is $\mathbb{R}^s$ so that $\phi(\cdot)$ may be expressed in the form $(\phi^{(1)}(\cdot), \phi^{(2)}(\cdot), \cdots, \phi^{(s)}(\cdot))^\top$. Suppose that the space containing $\phi^{(i)}(\cdot)$ is **separable**, that is, there exists basis functions $\{\beta_j\}_{j=1}^{m}$ (m can be infinity), such that $\forall i \in \{1, \cdots, s\}$,

$$\phi^{(i)}(x) = \sum_{j=1}^{m} \alpha_j^{(i)}(\phi)\beta_j(x).$$

On defining the matrix $M_\phi = \{\alpha_j^{(i)}(\phi)\}_{ij}$, we have

$$\phi(x) = M_\phi \cdot (\beta_1(x), \beta_2(x), \cdots, \beta_m(x))^\top$$

and let

$$\Gamma(x) = (\beta_1(x), \beta_2(x), \cdots, \beta_m(x))^\top.$$

Denote M_ϕ^- to be the MP-inverse of M_ϕ. Finally, for any $\psi_k \in \Psi_k$, define

$$I_i(\psi_k) = \{z : \psi_k^{(i)}(z) > \max_{j \neq i} \psi_k^{(j)}(z)\}.$$

With this decomposition we can now characterize invariant mappings.

Theorem 3.5. *For any $\varepsilon > 0$, if $\forall \psi_k \in \Psi_k$, $\cup_i I_i(\psi_k) = \mathbb{R}^s$, then $\phi \in \Phi$ satisfies Definition 3.2 iff*

$$\exists f \in Ker(M_\phi) \ s.t. \ \sum_{i=1}^{k} \mathbb{P}_{x \sim \tilde{\mathcal{D}}_i}(\Gamma(x) + f(x) \in M_\phi^- I_i(\psi_k)) \leqslant \varepsilon.$$

Above, the condition $\cup_i I_i(\psi_k) = \mathbb{R}^s$ is necessary to ensure that there will be no ties between the k weights produced by ψ_k. Note that our previous results do not require this condition.

4 Experiments

We assessed the performance of our approach on several datasets: (a) synthetic data based on those in biomedical studies [23], (b) colored MNIST [6], (c) PACS [14]. Our experiments confirm the conclusion (Section 3.1) that observing more domains improves generalization performance on an unseen one. For (a), we compared with logistic regression and random forest, whereas (b) and (c) were benchmarked against the state-of-the-art algorithms, IRM [2] and CIDDG [16]. Our code was adapted from the LAFTR code base [18], but the decoder was dropped to be constistent with our theoretical setting[1]. The prediction loss was taken to be binary cross-entropy.

[1]Extending the theory to include the decoder is an interesting direction for future work.

Synthetic Data. We consider synthetic data settings with $k = 4$ and $k = 10$. In each case, to sample a data point from a domain $\mathcal{D}_i$, a pair (x_j, y_j), $x_j \in \mathbb{R}^{30}, y_j \in \{0, 1\}$ is generated with $x_j \sim \mathcal{N}(\mu_i, \Sigma_i)$. The outcome y_j is generated so that a part of the relationship between x_j and y_j remains invariant across domains, while the other part varies. To operationalize this, we select a random subset $\mathcal{A}$ of covariates, a base rate b_i, and a set of functions $\{f_{\text{inv}}, f_1, \ldots, f_k\}$. We then sample $y_j \sim \text{Ber}(b_i)$ and accept (x_j, y_j) if $y_j = \mathbf{1}(f_{\text{inv}}(x_{j,\mathcal{A}}, \epsilon_{j,i}) > 0) = \mathbf{1}(f_i(x_{j,\mathcal{A}^c}) > 0)$, where $\epsilon_{j,i} \sim F_i$ is an additional small error term. Here, the parameters $\mu_i, \Sigma_i, b_i, F_i, f_i$ vary between the domains whereas f_{inv} and $\mathcal{A}$ remain invariant; $\epsilon_{j,i}$ ensures that the invariant signal between domains is not strong compared to the domain-specific one. Table 1 reports the performance of our algorithm on a new unseen domain of the same form as above, but with different parameters $\mu_{k+1}, \Sigma_{k+1}, b_{k+1}, F_{k+1}, f_{k+1}$. (Training involved 5000 samples from each seen domain.) Observe that the test accuracies increase from $k = 4$ to $k = 10$. We uniformly outperform both baselines by a notable margin.

Table 1: Test domain classification accuracy on (a) synthetic data where each of the functions $f_{\text{inv}}, f_1, \ldots, f_k, f_{k+1}$ contain a linear component and an interaction term; (b) similar synthetic data with responses generated differently (Section B, Supplementary) and each of the aforementioned functions now contain logical OR of two linear functions.

Algorithm	(a) 4-Domain	(a) 10-Domain	(b) 4-Domain	(b) 10-Domain
RVR	**90.6**%	**95.6**%	**86.1**%	**93.4**%
Logistic Regression	82.3%	86.2%	82.6%	86.7%
Random Forest	79.4%	89.0%	85.0%	88.3%

Colored MNIST. The colored MNIST data was generated from the MNIST database on handwritten digits [6]; here, the digit color acts as a spurious signal and the digit shape acts as the invariant signal. We perform binary classification on several versions of this dataset, following experimental setups similar to [2]. In Table 3, "A%-**shape** B%-**color**" refers to a setting with two training domains ($10,000$ samples each) both containing A% correlation between digit shape and labels: digits $0-4$ receive label 0 w.p. A% (1 o.w.), and digits $5-9$ receive label 1 w.p. A% (0 o.w.). In addition, there is a B% domain-specific correlation between digit color and labels: domain 1 (resp. domain 2) associates the color red with label 0 (resp. 1) and green with label 1 (resp. 0) w.p. B%. The last setting in Table 3 contains 6 training domains with shape-label correlation similar to that for the first, but the color-label correlation varies largely across domains, with each one consisting of mixtures of 2-3 different digit colors. The unseen test domains constitute either single color digits or a random mixture of red-green digit colors assigned independent of the label, and the same shape-label correlation as the corresponding training data. Our algorithm beats IRM and CIDDG in multiple settings, and performs comparably in others. Finally, Table 2 reports our performance when only 3 of the 6 domains from this setting are used as the training data (same test data). Once again, test accuracy improves remarkably with more seen domains.

Table 2: Multi-domain comparison of test accuracy on colored MNIST with 100% digit-label correlation and varying color-label correlations. The second column is the same as Table 3, row 3.

	3-Domain	6-Domain
RVR	86.1%	97.7%

Table 3: Test accuracy on several colored MNIST settings. *Target* denotes the digit color of the test domain. Details of the color-label correlation for row 3 can be found in Section B, Supplementary.

	Setting	Target	k	RVR	IRM	CIDDG
1.	100%-shape 90%-color	purple	2	**97.5**%	94.3%	95.7%
2.	75%-shape 80%-color	red-green	2	69.7*%	69.0%	**71.1**%
3.	100%-shape, unequal color	white	6	**97.7**%	94.7%	96.9%

PACS. To conclude, we examine our performance on PACS, an image-style dataset made of Photos, Art, Cartoon, and Sketch, which has been repeatedly used [15] to benchmark domain generalization algorithms. We specifically consider images labeled giraffes (label 0) or elephants (label 1), which leads to $384, 540, 803, 1493$ samples respectively. Each domain alternates as the target domain, while the algorithm trains on the rest. Once again, our algorithm beats (Table 4) both baselines across the board, and by a significant margin in most settings.

Table 4: Test accuracy on two types of images obtained from PACS

Target	RVR	IRM	CIDDG
P	**70.7**%	57.6%	62.1%
A	**66.7**%	64.2%	59.9%
C	**80.8**%	75.0%	73.8%
S	**54.3**%	54.0%	53.4%

5 Discussion

One natural question is whether we can extend the theoretical results to cross-entropy or other notions of loss. Next, our theoretical analysis does not yet fully capture our intuition regarding the conditions under which we believe our approach will succeed.

Our work forges a new path to address a major problem in biomedical research, where high-dimensional datasets are frequently encountered, and predictive algorithms increasingly used to inform personalized medical care. While strategies to ensure generalizability of these algorithms beyond the populations studied during training have been lacking, we are encouraged by our experimental results and have initiated engagement with the biomedical applications that inspired this work.

Acknowledgements

This work was supported in part by the Center for Research on Computation and Society (Harvard SEAS), the Harvard Data Science Initiative, NSF CCF-1763665, NIH/NCI 5T32CA009337-39, and NSF-DMS 1810829.

References

[1] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Domain generalization via invariant representation under domain-class dependency. 2018.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, May 2010.

[4] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in neural information processing systems*, pages 2178–2186, 2011.

[5] Peter Bühlmann. Invariance, causality and robustness. *arXiv preprint arXiv:1812.08233*, 2018.

[6] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[7] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 994–1003, 2018.

[8] Harrison Edwards and Amos Storkey. Censoring Representations with an Adversary. *arXiv.org*, November 2015.

[9] Sarah Erfani, Mahsa Baktashmotlagh, Masoud Moshtaghi, Vinh Nguyen, Christopher Leckie, James Bailey, and Ramamohanarao Kotagiri. Robust domain generalisation by enforcing distribution invariance. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 1455–1461. AAAI Press/International Joint Conferences on Artificial Intelligence, 2016.

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[11] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016.

[12] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015.

[13] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012.

[14] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[15] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018.

[16] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018.

[17] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The Variational Fair Autoencoder. *arXiv.org*, November 2015.

[18] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

[19] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. *arXiv preprint arXiv:1911.07661*, 2019.

[20] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[21] Li Niu, Wen Li, and Dong Xu. Multi-view domain generalization for visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4201, 2015.

[22] Li Niu, Wen Li, Dong Xu, and Jianfei Cai. An exemplar-based multi-view domain generalization framework for visual recognition. *IEEE transactions on neural networks and learning systems*, 29(2):259–272, 2016.

[23] Prasad Patil and Giovanni Parmigiani. Training replicable predictors in multiple studies. *Proceedings of the National Academy of Science USA*, March 2018.

[24] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[25] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

[26] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, pages 5334–5344, 2018.

[27] Ye Wang, Toshiaki Koike-Akino, and Deniz Erdogmus. Invariant representations from adversarially censored autoencoders. *arXiv preprint arXiv:1805.08097*, 2018.

[28] Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014.

[29] Richard Zemel, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Ma- chine Learning*, 2013.