EMPIRICAL OR INVARIANT RISK MINIMIZATION? A SAMPLE COMPLEXITY PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, invariant risk minimization (IRM) was proposed as a promising solution to address out-of-distribution (OOD) generalization. However, it is unclear when IRM should be preferred over the widely-employed empirical risk minimization (ERM) framework. In this work, we analyze both these frameworks from the perspective of sample complexity, thus taking a firm step towards answering this important question. We find that depending on the type of data generation mechanism, the two approaches might have very different finite sample and asymptotic behavior. For example, in the covariate shift setting we see that the two approaches not only arrive at the same asymptotic solution, but also have similar finite sample behavior with no clear winner. For other distribution shifts such as those involving confounders or anti-causal variables, however, the two approaches arrive at different asymptotic solutions where IRM is guaranteed to be *close* to the desired OOD solutions in the finite sample regime for polynomial generative models, while ERM is biased even asymptotically. We further investigate how different factors — the number of environments, complexity of the model, and IRM penalty weight — impact the sample complexity of IRM in relation to its distance from the OOD solutions.

1 INTRODUCTION

A recent study shows that models trained to detect COVID-19 from chest radiographs rely on spurious factors such as the source of the data rather than the lung pathology (DeGrave et al., 2020). This is just one of many alarming examples of spurious correlations failing to hold outside a specific training distribution. In one commonly cited example, Beery et al. (2018) trained a convolutional neural network (CNN) to classify camels from cows. In the training data, most pictures of the cows had green pastures, while most pictures of camels were in the desert. The CNN picked up the spurious correlation and associated green pastures with cows thus failing to classify cows on beaches.

Recently, Arjovsky et al. (2019) proposed a framework called invariant risk minimization (IRM) to address the problem of models inheriting spurious correlations. They showed that when data is gathered from multiple environments, one can learn to exploit invariant causal relationships, rather than relying on varying spurious relationships, thus learning robust predictors. More recent work suggests that empirical risk minimization (ERM) is still state-of-the-art on many problems requiring OOD generalization (Gulrajani & Lopez-Paz, 2020). This gives rise to a fundamental question: when is IRM better than ERM (and vice versa)? In this work, we seek to answer this question through a systematic comparison of the sample complexity of the two approaches under different types of train and test distributional mismatches.

The distribution shifts $\left(\mathbb{P}^{\mathsf{train}}(X,Y) \neq \mathbb{P}^{\mathsf{test}}(X,Y)\right)$ that we consider informally stated satisfy an *invariance condition* – there exists a representation Φ^* of the covariates such that $\mathbb{P}^{\mathsf{train}}(Y|\Phi^*(X)) = \mathbb{P}^{\mathsf{test}}(Y|\Phi^*(X)) = \mathbb{P}(Y|\Phi^*(X))$. A special case of this occurs when Φ^* is identity – $\mathbb{P}^{\mathsf{train}}(X) \neq \mathbb{P}^{\mathsf{test}}(X)$ but $\mathbb{P}^{\mathsf{train}}(Y|X) = \mathbb{P}^{\mathsf{test}}(Y|X)$ – such a shift is known as a *covariate-shift* (Gretton et al., 2009). In many other settings Φ^* may not be identity (denoted as I), examples include settings with confounders or anti-causal variables (Pearl, 2009) where covariates appear spuriously correlated with the label and $\mathbb{P}^{\mathsf{train}}(Y|X) \neq \mathbb{P}^{\mathsf{test}}(Y|X)$. We use causal Bayesian networks to illustrate these shifts in Figure 1. Suppose $X^e = [X_1^e, X_2^e]$ represents the image, where X_1^e is the shape of the animal and X_2^e is the background color, Y^e is the label of the animal, and e is the index of the

Table 1: Summary of (empirical) IRM vs. ERM for finite hypothesis class $\mathcal{H}_\Phi$. ϵ: slack in IRM constraints, ν: approximation w.r.t optimal risk, δ: failure probability, $\mathcal{E}_{tr}$: set of training environments, n: data dimension, p: degree of the generative polynomial, L, L': bound on loss & its gradients.

Assumptions	Method	Sample complexity	OOD
Covariate shift case: $\mathbb{E}^e[Y^e \vert X^e]$ is invariant (Proposition 4)	ERM	$\frac{8L^2}{\nu^2}\log\left(\frac{2\vert\mathcal{H}_\Phi\vert}{\delta}\right)$	Yes
	IRM	$\max\left\{\frac{8L^2}{\nu^2}\log\left(\frac{4\vert\mathcal{H}_\Phi\vert}{\delta}\right), \frac{16L'^4}{\epsilon^2}\log\left(\frac{2}{\delta}\right)\right\}$	Yes
Confounder/Anti-causal variable case: $\mathbb{E}^e[Y^e \vert \Phi^*(X^e)]$ is invariant, Linear, Polynomial models, $\vert\mathcal{E}_{tr}\vert = \mathcal{O}(n^p)$ (Proposition 5, 6,17)	ERM	$\frac{8L^2}{\nu^2}\log\left(\frac{2\vert\mathcal{H}_\Phi\vert}{\delta}\right)$	No
	IRM	$\frac{16L'^4}{\epsilon^2}\log\left(\frac{2\vert\mathcal{H}_\Phi\vert}{\delta}\right)$	Yes

environment/domain. In Figure 1a) X_2^e is independent of (Y^e, X_1^e), it represents the covariate shift case ($\Phi^* = \mathsf{I}$). In Figure 1b) X_2^e is spuriously correlated with Y^e through the confounder ε^e. In Figure 1c) X_2^e is spuriously correlated with Y^e as it is anti-causally related to Y^e. In both Figure 1b) and c) $\Phi^* \neq \mathsf{I}$; Φ^* is a block diagonal matrix that selects X_1^e.

Our setup assumes we are given data from multiple training environments satisfying the invariance condition, i.e., $\mathbb{P}(Y|\Phi^*(X))$ is the same across all of them. Ideally, we want to learn and predict using $\mathbb{E}[Y|\Phi^*(X)]$; this predictor has a desirable OOD behavior as we show later where we prove min-max optimality with respect to (w.r.t.) unseen test distributions satisfying the invariance condition. Our goal is to analyze and compare ERM and IRM's ability to learn $\mathbb{E}[Y|\Phi^*(X)]$ from finite training data acquired from a fixed number of training environments. Our analysis has two parts.

1) Covariate shift case ($\Phi^* = \mathsf{I}$): ERM and IRM achieve the same asymptotic solution $\mathbb{E}[Y|X]$. We prove (Proposition 4) that the sample complexity for both the methods is similar thus there is no clear winner between the two in the finite sample regime. For the setup in Figure 1a), both ERM and IRM get close to a model that only uses X_1^e.

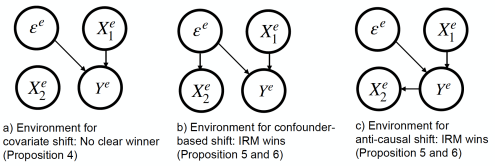

Figure 1: Causal Bayesian networks illustrating environments under different distribution shifts.

2) Confounder/Anti-causal variable case ($\Phi^* \neq \mathsf{I}$): We consider a family of structural equation models (linear and polynomial) that may contain confounders and/or anti-causal variables. For the class of models we consider, the asymptotic solution of ERM is biased and not equal to the desired $\mathbb{E}[Y|\Phi^*(X)]$. We prove that IRM can learn a solution that is within $\mathcal{O}(\sqrt{\epsilon})$ distance from $\mathbb{E}[Y|\Phi^*(X)]$ with a sample complexity that increases as $\mathcal{O}(\frac{1}{\epsilon^2})$ and increases polynomially in the complexity of the model class (Proposition 5, 6); ϵ (defined later) is the slack in IRM constraints. For the setup in Figure 1b) and c), IRM gets close to only using X_1^e, while ERM even with infinite data (Proposition 17 in the supplement) continues to use X_2^e. We summarize the results in Table 1.

Arjovsky et al. (2019) proposed the colored MNIST (CMNIST) dataset; comparisons on it showed how ERM-based models exploit spurious factors (background color). The CMNIST dataset relied on anti-causal variables. Many supervised learning datasets may not contain anti-causal variables (e.g. human labeled images). Therefore, we propose and analyze three new variants of CMNIST in addition to the original one that map to different real-world settings: i) covariate shift based CMNIST (CS-CMNIST): relies on selection bias to induce spurious correlations, ii) confounded CMNIST (CF-CMNIST): relies on confounders to induce spurious correlations, iii) anti-causal CMNIST (AC-CMNIST): this is the original CMNIST proposed by Arjovsky et al. (2019), and iv) anti-causal and confounded (hybrid) CMNIST (HB-CMNIST): relies on confounders and anti-causal variables to induce spurious correlations. On the latter three datasets, which belong to the $\Phi^* \neq \mathsf{I}$ class described above, IRM shows a much better OOD behavior than ERM, which performs poorly regardless of the data size. However, we find IRM and ERM have a similar performance on CS-CMNIST ($\Phi^* = \mathsf{I}$) with no clear winner. These results are consistent with our theory and are further validated with experiments in the regression setting.

2 RELATED WORKS

IRM based works. Following the original work IRM from Arjovsky et al. (2019), there have been several interesting works — (Teney et al., 2020; Krueger et al., 2020; Ahuja et al., 2020; Chang et al., 2020; Koyama & Yamaguchi, 2020; Mahajan et al., 2020) is an incomplete representative list — that build new methods inpired from IRM to address the OOD generalization problem. Arjovsky et al. (2019) prove OOD guarantees for linear models with access to infinite data from finite environments. We generalize these results in several ways. We provide a first finite sample analysis of IRM. We characterize the impact of model complexity, number of environments, weight of IRM penalty on the sample complexity and its distance from the OOD solution for linear and polynomial models.

Theory of domain generalization and domain adaption. Following the seminal works (Ben-David et al., 2007; 2010), there have been many interesting works — (Muandet et al., 2013; Ajakan et al., 2014; Zhao et al., 2019; Albuquerque et al., 2019; Li et al., 2017; Piratla et al., 2020; Matsuura & Harada, 2020; Deng et al., 2020; David et al., 2010; Pagnoni et al., 2018) is an incomplete representative list (see Redko et al. (2019) for further references) — that build the theory of domain adaptation and generalization and construct new methods based on it. While many of these works develop bounds on loss over the target domain using train data and unlabeled target data, some (Ben-David & Urner, 2012; David et al., 2010; Pagnoni et al., 2018) analyze the finite sample (PAC) guarantees for domain adaptation under covariate shifts. These works (Ben-David & Urner, 2012; David et al., 2010; Pagnoni et al., 2018) access unlabeled data from a target domain, which we do not. Instead, we have data from multiple training domains (as in domain generalization). In these works, the guarantees are w.r.t. a specific target domain, while we provide (for linear and polynomial models) worst-case guarantees w.r.t. all the unseen domains satisfying the invariance condition. Also, we consider a larger family of distribution shifts including covariate shifts.

3 SAMPLE COMPLEXITY OF INVARIANT RISK MINIMIZATION

3.1 INVARIANT RISK MINIMIZATION

We start with some background on IRM (Arjovsky et al., 2019). Consider a dataset $D = \{D^e\}_{e\in\mathcal{E}_{tr}}$, which is a collection of datasets $D^e = \{(x_i^e, y_i^e, e)\}_{i=1}^{n_e}\}$ obtained from a set of training environments $\mathcal{E}_{tr}$, where e is the index of the environment, i is the index of the data point in the environment, n_e is the number of points from environment, $x_i^e \in \mathcal{X} \subseteq \mathbb{R}^n$ is the feature value and $y_i^e \in \mathcal{Y} \subseteq \mathbb{R}$ is the corresponding label. Define a probability distribution $\{\pi^e\}_{e\in\mathcal{E}_{tr}}$, π^e is the probability that a training data point is from environment e. Define a probability distribution of points conditional on environment e as $\mathbb{P}^e$, $(X^e, Y^e) \sim \mathbb{P}^e$. Define the joint distribution $\bar{\mathbb{P}}$, $(X^e, Y^e, e) \sim \bar{\mathbb{P}}, d\bar{\mathbb{P}}(X^e, Y^e, e) = \pi^e d\mathbb{P}^e(X^e, Y^e)$. D is a collection of i.i.d. samples from $\bar{\mathbb{P}}$. Define a predictor $f : \mathcal{X} \to \mathbb{R}$ and the space $\mathcal{F}$ of all the possible maps from $\mathcal{X} \to \mathbb{R}$. Define the risk achieved by f in environment e as $R^e(f) = \mathbb{E}^e\big[\ell\big(f(X^e), Y^e\big)\big]$, where ℓ is the loss, $f(X^e)$ is the predicted value, Y^e is the corresponding label and $\mathbb{E}^e$ is the expectation conditional on environment e. The overall expected risk across the training environments is $R(f) = \sum_{e\in\mathcal{E}_{tr}} \pi^e R^e(f)$. We are interested in two settings: regression (square loss) and binary-classification (cross-entropy loss). In the main body, our focus is regression (square loss) and we mention wherever the results extend to binary-classification (cross-entropy). We discuss these extensions in the supplement.

OOD generalization problem. We want to construct a predictor f that performs well across many unseen environments $\mathcal{E}_{all}$, where $\mathcal{E}_{all} \supseteq \mathcal{E}_{tr}$. For $o \in \mathcal{E}_{all}\backslash\mathcal{E}_{tr}$, the distribution $\mathbb{P}^o$ can be very different from the train environments. Next we state the OOD problem.

$$\min_{f\in\mathcal{F}} \max_{e\in\mathcal{E}_{all}} R^e(f) \tag{1}$$

The above problem is very challenging to solve since we only have access to data from training environments $\mathcal{E}_{tr}$ but are required to find the robust solution over all environments $\mathcal{E}_{all}$. Next, we make assumptions on $\mathcal{E}_{all}$ and characterize the optimal solution to equation 1.

Assumption 1. ***Invariance condition.*** *There exists a representation Φ^* that transforms X^e to $Z^e = \Phi^*(X^e)$ and $\forall e, o \in \mathcal{E}_{all}, \forall z \in \Phi^*(\mathcal{X})$ satisfies $\mathbb{E}^e[Y^e|Z^e = z] = \mathbb{E}^o[Y^o|Z^o = z]$. Also, $\forall e \in \mathcal{E}_{all}, \forall z \in \Phi^*(\mathcal{X})$, $\mathsf{Var}^e[Y^e|Z^e = z] = \xi^2$, where Var^e is the conditional variance.*

The above assumption is inspired from causality (Pearl, 2009). Φ^* acts as the causal feature extractor and from the definition of causal features, it follows that $\mathbb{E}^e[Y^e|Z^e = z]$ does not vary across

environments. When a human labels a cow she uses Φ^* to extract causal features from the pixels to identify cow while ignoring the background. The first part of the above assumption encompasses a large class of distribution shifts including standard covariate shifts (Gretton et al., 2009). Covariate shift assumes $\forall e, o \in \mathcal{E}_{all}, \forall x \in \mathcal{X}$, $\mathbb{P}(Y^e|X^e = x)$ and $\mathbb{P}(Y^o|X^o = x)$ are equal thus implying $\mathbb{E}^e[Y^e|X^e = x] = \mathbb{E}^o[Y^o|X^o = x]$. Therefore, for covariate shifts, Φ^* is identity in Assumption 1. A simple instance illustrating Assumption 1 with $\Phi^* = \mathsf{I}$ is when $Y^e \leftarrow g(X^e) + \varepsilon^e$, where $\mathbb{E}^e[\varepsilon^e] = 0, \mathbb{E}^e[(\varepsilon^e)^2] = \sigma^2$, $\varepsilon^e \perp X^e$. Using Assumption 1, we define the invariant map $m : \Phi^*(\mathcal{X}) \to \mathbb{R}$ as follows

$$\forall z \in \Phi^*(\mathcal{X}),\ m(z) = \mathbb{E}^e[Y^e|Z^e = z],\ \text{where } Z^e = \Phi^*(X^e) \tag{2}$$

Assumption 2. ***Existence of an environment where the invariant representation is sufficient.*** *$\exists$ an environment $e \in \mathcal{E}_{all}$ such that $Y^e \perp X^e|Z^e$*

Assumption 2 states there exists an environment where the information that X^e has about Y^e is also contained in Z^e. Define a composition $m \circ \Phi^*$, $\forall x \in \mathcal{X}$, $m \circ \Phi^*(x) = \mathbb{E}^e[Y^e|Z^e = \Phi^*(x)]$.

Proposition 1. *If ℓ is the square loss, and Assumptions 1 and 2 hold, then $m \circ \Phi^*$ solves the OOD problem (equation 1).*

The proofs of all the propositions are in the supplement. A similar result holds for the cross-entropy loss (discussion in supplement). For the rest of the paper, we focus on learning $m \circ \Phi^*$ as it solves the OOD problem. For covariate shifts $\Phi^* = \mathsf{I}$, $m(x) = \mathbb{E}^e[Y^e|X^e = x]$ is the OOD solution. In Arjovsky et al. (2019), a proof connecting $m \circ \Phi^*$ and OOD was not stated. Recently, in Koyama & Yamaguchi (2020), a result similar to Proposition 1 was shown but with a few differences. The authors assume conditional probabilities are invariant unlike our assumption that only requires conditional expectations and variances to be invariant. However, their result applies to more losses. $m \circ \Phi^*$ is the target we want to learn. Arjovsky et al. (2019) proposed IRM since standard min-max optimization over the training environments $\mathcal{E}_{tr}$ and ERM fail to learn $m \circ \Phi^*$ in many cases. The authors in Arjovsky et al. (2019) identify a crucial property of $m \circ \Phi^*$ and use it to define an object called invariant predictor that we define next.

Invariant predictor and IRM optimization. Define a representation map $\Phi : \mathcal{X} \to \mathcal{Z}$ from feature space to representation space $\mathcal{Z} \subseteq \mathbb{R}^q$. Define a classifier map, $w : \mathcal{Z} \to \mathbb{R}$ from representation space to real values. Define $\mathcal{H}_\Phi$ and $\mathcal{H}_w$ as the spaces of representations and classifiers respectively. A data representation Φ elicits an invariant predictor $w \circ \Phi$ across environments $e \in \mathcal{E}_{tr}$ if there is a classifier w that achieves the minimum risk simultaneously for all the environments, i.e., $\forall e \in \mathcal{E}_{tr}, w \in \arg\min_{\bar{w} \in \mathcal{H}_w} R^e(\bar{w} \circ \Phi)$. Observe that if we we transform the data with representation Φ^* then m will achieve the minimum risk simultaneously in all the environments. If $\Phi^* \in \mathcal{H}_\Phi$ and $m \in \mathcal{H}_w$, then $m \circ \Phi^*$ is an invariant predictor. IRM selects the invariant predictor with least sum risk across environments (results presented later can be adapted if invariant predictor was selected based on the worst-case risk over the environments as well) as follows:

$$\begin{aligned} &\min_{\Phi \in \mathcal{H}_\Phi, w \in \mathcal{H}_w} R(w \circ \Phi) = \sum_{e \in \mathcal{E}_{tr}} \pi^e R^e(w \circ \Phi) \\ &\text{s.t. } w \in \arg\min_{\bar{w} \in \mathcal{H}_w} R^e(\bar{w} \circ \Phi),\ \forall e \in \mathcal{E}_{tr} \end{aligned} \tag{3}$$

From the above discussion we know $m \circ \Phi^*$ is a feasible solution to equation 3. It is also the ideal solution we want IRM to find since it solves equation 1. Later in Propositions 4, 5, and 6, we show that IRM actually solves equation equation 1. For the setups in Proposition 5, and 6, conventional ERM based approaches fail thus justifying the need for above formulation.

3.2 Sample Complexity of Gradient Constraint Formulation of IRM

In Arjovsky et al. (2019), a gradient constrained alternate (derived below in equation 4) to equation 3 was proposed, which focuses on linear and scalar classifiers ($\mathcal{Z} = \mathbb{R}$, $\Phi : \mathcal{X} \to \mathbb{R}$, $\mathcal{H}_w = \mathbb{R}$). In this case, the composite predictor $w \circ \Phi$ is a multiplication of w and Φ written as $w \cdot \Phi$. (For binary-classification predictor's output $w \cdot \Phi(x)$ represents logits.) From the definition of invariant predictors and $\mathcal{H}_w = \mathbb{R}$ it follows that if $\forall \bar{w} \in \mathbb{R},\ R^e(1 \cdot \Phi) \leq R^e(\bar{w} \cdot \Phi)$, then Φ is an invariant predictor. For square and cross-entropy losses, $R^e(w \cdot \Phi)$ is convex in w. Therefore, a gradient constraint $\|\nabla_{w|w=1.0} R^e(w \cdot \Phi)\| = 0$ is equivalent to the condition that $\forall \bar{w} \in \mathbb{R},\ R^e(1 \cdot \Phi) \leq R^e(\bar{w} \cdot \Phi)$, which implies Φ is an invariant predictor. Recall that IRM aims is to search among invariant predictors and find one that minimizes the risk. We state this as a gradient constrained optimization as follows

$$\begin{aligned} &\min_{\Phi \in \mathcal{H}_\Phi} R(\Phi) \\ &\text{s.t. } \left\| \nabla_{w|w=1.0} R^e(w \cdot \Phi) \right\| = 0, \ \forall e \in \mathcal{E}_{tr} \end{aligned} \tag{4}$$

We propose an ϵ approximation of the above with ϵ slack in the constraint. Define $R^{'}(\Phi) = \sum_{e \in \mathcal{E}_{tr}} \pi^e \|\nabla_{w|w=1.0} R^e(w \cdot \Phi)\|^2$ and a set $\mathcal{S}^{\mathsf{IV}}(\epsilon) = \{\Phi \mid R^{'}(\Phi) \leq \epsilon, \Phi \in \mathcal{H}_\Phi\}$. Note that $R^{'}$ is very similar to the penalty defined in Arjovsky et al. (2019). The ϵ approximation of equation 4 is

$$\min_{\Phi \in \mathcal{S}^{\mathsf{IV}}(\epsilon)} R(\Phi) \tag{5}$$

If $\epsilon = 0$, then equation 4 and equation 5 are equivalent. In all the optimizations so far, the expectations are computed w.r.t. the distributions $\mathbb{P}^e$, which are unknown. Therefore, we develop an empirical version of equation 5 below (in equation 6) and call it empirical IRM (EIRM). We replace R and $R^{'}$ with empirical estimators $\hat{R}$ and $\hat{R}^{'}$ respectively. For R we use a simple plugin estimator (sample mean of loss across all the samples in D). For $R^{'}$ we construct a new estimator that enables the use of standard concentration inequalities. Define a set $\hat{\mathcal{S}}^{\mathsf{IV}}(\epsilon) = \{\Phi \mid \hat{R}^{'}(\Phi) \leq \epsilon, \Phi \in \mathcal{H}_\Phi\}$.

$$\min_{\Phi \in \hat{\mathcal{S}}^{\mathsf{IV}}(\epsilon)} \hat{R}(\Phi) \tag{6}$$

If we replaced $\hat{\mathcal{S}}^{\mathsf{IV}}(\epsilon)$ with $\mathcal{H}_\Phi$ in equation 6, then we get the standard ERM. ERM aims to solve $\min_{\Phi \in \mathcal{H}_\Phi} \hat{R}(\Phi)$. The sample complexity analysis of ERM aims to understand the distance between the empirical solutions and the expected solutions as a function of the number of samples. Similarly, we seek to understand the relationship between solutions of equation 6 and equation 5.

Assumption 3. ***Bounded loss and bounded gradient of the loss.*** *$\exists\ L < \infty$, $L' < \infty$ such that $\forall \Phi \in \mathcal{H}_\Phi, \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, |\ell(\Phi(x), y)| \leq L, |\frac{\partial \ell(w \cdot \Phi(x), y)}{\partial w}|_{w=1.0}| \leq L^{'}$.*

If every Φ in the hypothesis class $\mathcal{H}_\Phi$ is bounded by M and the label space $\mathcal{Y}$ is bounded, then for both square and cross-entropy loss, $\ell(\Phi(\cdot), \cdot)$ and $\frac{\partial \ell(w \cdot \Phi(\cdot), \cdot)}{\partial w}|_{w=1.0}$ are bounded. Define $\kappa = \min_{\Phi \in \mathcal{H}_\Phi} |R^{'}(\Phi) - \epsilon|$; κ measures how close any penalty can get to the boundary ϵ. κ quantifies how good the finite sample approximation $\hat{R}^{'}$ need to be in order to get $\hat{\mathcal{S}}^{\mathsf{IV}}(\epsilon) = \mathcal{S}^{\mathsf{IV}}(\epsilon)$. Define ν to quantify the approximation w.r.t. optimal risk.

Proposition 2. *For every $\nu > 0, \epsilon > 0$ and $\delta \in (0, 1)$, if $\mathcal{H}_\Phi$ is a finite hypothesis class, Assumption 3 holds, $\kappa > 0$, and if the number of samples $|D|$ is greater than $\max\left\{\frac{16L'^4}{\kappa^2}, \frac{8L^2}{\nu^2}\right\} \log\left(\frac{4|\mathcal{H}_\Phi|}{\delta}\right)$, then with a probability at least $1 - \delta$, every solution $\hat{\Phi}$ to EIRM (equation 6) is a ν approximation of IRM, i.e. $\hat{\Phi} \in \mathcal{S}^{\mathsf{IV}}(\epsilon)$, $R(\Phi^*) \leq R(\hat{\Phi}) \leq R(\Phi^*) + \nu$, where Φ^* is a solution to IRM (equation 5).*

Proof Sketch. The standard analysis in learning theory on ERM or regularized/constrained ERM typically relies on linearly separable loss functions. In such cases, we can use standard plug-in estimators and analyze their behavior using concentration inequalities. In our setting, $R^{'}$ is a weighted sum of squares of expectation and thus it is not linearly separable. We develop a new way of expressing $R^{'}$ that allows us to make it linearly separable. Next, in order to ensure $R(\Phi^*) \leq R(\hat{\Phi}) \leq R(\Phi^*) + \nu$ we first need to guarantee that the set of invariant predictors is exactly recovered, i.e., $\hat{\mathcal{S}}^{\mathsf{IV}}(\epsilon) = \mathcal{S}^{\mathsf{IV}}(\epsilon)$ (exact recovery is typically not required in existing constrained analysis such as Woodworth et al. (2017) Agarwal et al. (2018)). We show that if the number of samples grow as $\frac{1}{\kappa^2}$ even the closest points on either side of the boundary of the set $\mathcal{S}^{\mathsf{IV}}(\epsilon)$ are correctly discriminated, which guarantees exact recovery of $\mathcal{S}^{\mathsf{IV}}(\epsilon)$. Once the exact set is recovered, beyond this we use standard learning theory tools to ensure $R(\Phi^*) \leq R(\hat{\Phi}) \leq R(\Phi^*) + \nu$

The above result holds for both square and cross-entropy loss. For ease of exposition, we use the standard setting of finite hypothesis class and extend all the results to infinite hypothesis classes in the supplement (summary of insights from the extension are in Section 3.3.2). Next, we state a standard result on ERM's sample complexity. Define a Φ^+ such that $\Phi^+ \in \arg\min_{\Phi \in \mathcal{H}_\Phi} R(\Phi)$

Proposition 3. *(Shalev-Shwartz & Ben-David, 2014) For every $\nu > 0$ and $\delta \in (0, 1)$, if $\mathcal{H}_\Phi$ is a finite hypothesis class, Assumption 3 holds, and if the number of samples $|D|$ is greater than $\frac{8L^2}{\nu^2} \log\left(\frac{2|\mathcal{H}_\Phi|}{\delta}\right)$, then with a probability at least $1 - \delta$, every solution $\Phi^\dagger$ to ERM is an ν approximation of expected risk minimization, i.e., $R(\Phi^+) \leq R(\Phi^\dagger) \leq R(\Phi^+) + \nu$.*

Proposition 2 vs. 3 Since $\kappa \leq \epsilon$, the sample complexity of EIRM grows at least as $\mathcal{O}(\max\{\frac{1}{\epsilon^2}, \frac{1}{\nu^2}\})$. Let us look at the two terms inside $\max$- i) $\frac{1}{\nu^2}$ growth term is similar to ERM, it ensures ν approximate optimality in the overall risk R, ii) $\frac{1}{\epsilon^2}$ growth ensures the IRM penalty $R^{'}$ is less than ϵ. A direct comparison of sample complexities in Propositions 2 and 3 suggests that the sample complexity of EIRM is higher than ERM, which is not the complete picture. The two approaches may not converge to the same solutions and IRM may converge to a solution with better OOD behavior than one achieved by ERM. Therefore, a fair comparison is only possible when we also study the OOD properties of the solutions achieved by the two approaches, which is the subject of the next section.

3.3 OOD Performance: ERM vs. IRM

We divide the comparisons based on distributional shift assumptions that decide whether ERM and IRM arrive at the same asymptotic solutions or not.

3.3.1 Covariate shift

Assumption 4. ***Invariance w.r.t. all the features.*** *$\forall e, o \in \mathcal{E}_{all}$ and $\forall x \in \mathcal{X}$, $\mathbb{E}[Y^e|X^e = x] = \mathbb{E}[Y^o|X^o = x]$. $\forall e \in \mathcal{E}_{all}$, $X^e \sim \mathbb{P}^e_{X^e}$ and the support of $\mathbb{P}^e_{X^e}$ is equal to $\mathcal{X}$.*

As stated earlier, the first part of the above assumption follows from standard covariate shift assumptions (Gretton et al., 2009) and is a special case of the first part of the Assumption 1 with Φ^* set to I. If $\Phi^* = \mathsf{I}$, then m (equation 2), which simplifies to $m(x) = \mathbb{E}[Y^e|X^e = x]$, solves the OOD problem equation 1. A generative model that satisfies the above Assumption 4 is given as

$$Y^e \leftarrow g(X^e) + \varepsilon^e,\ \mathbb{E}[\varepsilon^e] = 0,\ \varepsilon^e \perp X^e,\ \mathbb{E}[(\varepsilon^e)^2] = \sigma^2 \tag{7}$$

In the above model X^e is the cause, Y^e is the effect, and g a general non-linear function (it satisfies Assumption 4 with $m = g$). Next, we compare ERM and IRM's ability to learn m under covariate shifts. In Figure 1 a), we show Define $\tilde{\kappa} = \min_{\Phi_1, \Phi_2 \in \mathcal{H}_\Phi, \Phi_1 \neq \Phi_2} |R(\Phi_1) - R(\Phi_2)|$, which measures the minimum separation between the risks of any two distinct hypothesis in $\mathcal{H}_\Phi$.

Proposition 4. *Let ℓ be the square loss. For every $\nu > 0, \epsilon > 0$ and $\delta \in (0, 1)$, if $\mathcal{H}_\Phi$ is a finite hypothesis class, $m \in \mathcal{H}_\Phi$, Assumptions 3, 4 hold, and*

- *if the number of samples $|D|$ is greater than $\max\left\{\frac{8L^2}{\nu^2}\log(\frac{4|\mathcal{H}_\Phi|}{\delta}), \frac{16L'^4}{\epsilon^2}\log(\frac{2}{\delta})\right\}$, then with a probability at least $1-\delta$, every solution $\hat{\Phi}$ to EIRM (equation 6) satisfies $R(m) \leq R(\hat{\Phi}) \leq R(m) + \nu$. If also $\nu < \tilde{\kappa}$, then $\hat{\Phi} = m$.*

- *if the number of samples $|D|$ is greater than $\frac{8L^2}{\nu^2}\log(\frac{2|\mathcal{H}_\Phi|}{\delta})$, then with a probability at least $1-\delta$, every solution $\Phi^\dagger$ to ERM satisfies $R(m) \leq R(\Phi^\dagger) \leq R(m) + \nu$. If also $\nu < \tilde{\kappa}$, then $\Phi^\dagger = m$.*

Implications of Proposition 4. ERM and EIRM both asymptotically achieve the ideal OOD solution; the above proposition helps compare them in a finite sample regime. The second term inside the $\max$ for EIRM, $\frac{16L'^4}{\epsilon^2}\log(\frac{2}{\delta})$, does not depend on the size of the hypothesis class. Hence, for large hypothesis classes, the sample complexity of EIRM equals $\frac{8L^2}{\nu^2}\log(\frac{4|\mathcal{H}_\Phi|}{\delta})$. Consequently, the sample complexity of EIRM and ERM differs by a constant $\frac{8L^2}{\nu^2}\log(2)$. Thus we conclude, for large hypothesis classes, both ERM and EIRM have similar sample complexity. Next, we contrast the sample complexity of EIRM in Proposition 4, $\mathcal{O}(\frac{1}{\nu^2})$, to Proposition 2, $\mathcal{O}(\max\{\frac{1}{\epsilon^2}, \frac{1}{\nu^2}\})$; the additional covariate shift assumption in Proposition 4 helps get to a lower sample complexity of $\mathcal{O}(\frac{1}{\nu^2})$. In Proposition 4, we assumed square loss, but a similar result extends to cross-entropy loss as well.

3.3.2 Distributional Shift with Confounders and (or) Anti-Causal Variables

In this section, we consider more general models than equation 7, which only contained cause X^e and effect Y^e. We also allow confounders and anti-causal variables. However, we restrict g to polynomials. We start with linear models from Arjovsky et al. (2019).

Assumption 5.

$$\begin{aligned} &e \sim \mathsf{Categorical}(\{\pi^o\}_{o \in \mathcal{E}_{tr}}),\ \forall o \in \mathcal{E}_{tr}, \pi^o > 0 \\ &Y^e \leftarrow \gamma^\mathsf{T}(Z^e_1) + \varepsilon^e,\ \varepsilon^e \perp Z^e_1,\ \mathbb{E}[\varepsilon^e] = 0,\ \mathbb{E}[(\varepsilon^e)^2] = \sigma^2,\ |\varepsilon^e| \leq \varepsilon^{\mathsf{sup}} \\ &X^e \leftarrow S(Z^e_1, Z^e_2) \end{aligned} \tag{8}$$

We assume that Z_1 component of S is invertible, i.e. $\exists\ \tilde{S}$ such that $\tilde{S}(S(Z_1, Z_2)) = Z_1$, and $\gamma \neq 0$. $\forall e \in \mathcal{E}_{tr}, \pi^e \geq \frac{\pi^{\min}}{|\mathcal{E}_{tr}|} > 0$. Define $\Sigma^e = \mathbb{E}[X^e X^{e,\mathsf{T}}]$. $\forall e \in \mathcal{E}_{tr}$, Σ^e is positive definite. The support of distribution of $Z^e = (Z_1^e, Z_2^e)$, $\mathbb{P}^e_{Z^e}$, is bounded and the norm of S, $\|S\| = \sigma_{\max}(S)$ (maximum singular value of S), is also bounded.

In the above model, Z_1^e is the cause of X^e and Y^e but may not be directly observed. Z_2^e may be arbitrarily correlated with Z_1^e and ϵ^e. We observe a scrambled transformation of (Z_1^e, Z_2^e) in X^e. If Z_2^e is an effect of Y^e ($Z_2^e \leftarrow Y^e + N^e$), then Z_2^e is an anti-causal variable. If H^e causes both ε^e ($\varepsilon^e \leftarrow H^e + \bar{N}^e$) and Z_2^e ($Z_2^e \leftarrow H^e + \tilde{N}^e$), then H^e is a confounder. In both these cases, Z_2^e is spuriously correlated with the label Y^e. Consequently, the standard ERM based models estimate Z_2^e from X^e and end up being biased w.r.t the desired OOD model, which does not use Z_2^e. If Assumptions 5, 2 hold, then a linear model $\tilde{S}^\mathsf{T}\gamma$ ($X^e \xrightarrow{\text{predict}} \gamma^\mathsf{T}\tilde{S}X^e = \gamma^\mathsf{T} Z_1^e$) solves the OOD problem in equation 1 ($\Phi^* = \tilde{S}$, $m = \gamma^\mathsf{T}$ in Proposition 1) and it relies only on Z_1^e. In the supplement (Proposition 17), we prove that ERM based models do not recover $\tilde{S}^\mathsf{T}\gamma$.

Assumption 6. ***Linear general position.*** *A set of training environments $\mathcal{E}_{tr}$ is said to lie in a linear general position of degree r for some $r \in \mathbb{N}$ if $|\mathcal{E}_{tr}| > n - r + n/r$ and for all non-zero $x \in \mathbb{R}^n$*

$$\mathsf{dim}\Big(\mathsf{span}\Big\{\mathbb{E}^e[X^e X^{e,\mathsf{T}}]x - \mathbb{E}^e[X^e \varepsilon^e]\Big\}_{e \in \mathcal{E}_{tr}}\Big) > n - r \tag{9}$$

where span *is the linear span,* dim *is the dimension, and recall n is dimension of X^e. This assumption checks for diversity in the environments and holds almost everywhere (Arjovsky et al., 2019).*

Assumption 7. ***Inductive bias.*** *$\mathcal{H}_\Phi$ is a finite set of linear models parametrized by $\Phi \in \mathbb{R}^n$ (output $\Phi^\mathsf{T} X^e$). $\tilde{S}^\mathsf{T}\gamma \in \mathcal{H}_\Phi$. $\exists\ \omega > 0, \Omega > 0$, s.t. $\forall \Phi \in \mathcal{H}_\Phi,\ \omega \leq \|\Phi\|^2 \leq \Omega$ & $2\omega \leq \|\tilde{S}^\mathsf{T}\gamma\|^2 \leq \frac{2}{3+2\sqrt{2}}\Omega$*

Informally stated, the above assumption requires the OOD optimal predictor $\tilde{\mathcal{S}}^\mathsf{T}\gamma$ to lie in the interior of the search space and not on the boundary. If Assumptions 5, 7 hold, then Assumption 3 holds. Hence, we can use the bounds L and L' on ℓ and $\frac{\partial \ell(w \cdot \Phi(\cdot), \cdot)}{\partial w}|_{w=1.0}$ respectively in our next result. Define the minimum eigenvalue across all Σ^e as $\lambda_{\min} = \min_{e \in \mathcal{E}_{tr}} \lambda_{\min}(\Sigma_e)$, $\epsilon_{\mathsf{th}} = \frac{(24-16\sqrt{2})}{3}\frac{\pi^{\min}}{|\mathcal{E}_{tr}|}(\omega\lambda_{\min})^2$ and $\tau = \frac{1}{2\omega\lambda_{\min}}\sqrt{\frac{3|\mathcal{E}_{tr}|}{2\pi^{\min}}}$. Next, we analyze how EIRM learns $\tilde{\mathcal{S}}^\mathsf{T}\gamma$.

Proposition 5. *Let ℓ be the square loss. For every $\epsilon \in (0, \epsilon_{\mathsf{th}})$ and $\delta \in (0,1)$, if Assumptions 5, 6 (with $r = 1$), 7 hold and if the number of data points $|D|$ is greater than $\frac{16L'^4}{\epsilon^2}\log\big(\frac{2|\mathcal{H}_\Phi|}{\delta}\big)$, then with a probability at least $1 - \delta$, every solution $\hat{\Phi}$ to EIRM (equation 6) satisfies $\hat{\Phi} = (\tilde{S}^\mathsf{T}\gamma)\alpha$, where $\alpha \in [\frac{1}{1+\tau\sqrt{\epsilon}}, \frac{1}{1-\tau\sqrt{\epsilon}}]$.*

Proof Sketch. In learning theory it is common to analyze the concentration of empirical risks around the expected risks. In our case, we have a target ideal solution to equation 1 ($\tilde{S}^\mathsf{T}\gamma$) and we want our empirical solutions to concentrate around that. A direct finite sample approximation of equation 4 is hard to analyze. Therefore, we introduce an intermediate problem in equation 5 and then develop a finite sample approximation of it in equation 6. We first show that solving equation 5 leads to solutions in the neighborhood of the target. To show this we use the linear general position assumption. Next, we connect equation 6 and equation 5 using our new estimator for $R^{'}$ and Hoeffding's inequality.

Implications of Proposition 5. 1. Convergence rate of ERM vs. EIRM: Recall that ϵ is the slack on IRM penalty $R^{'}$. If ϵ is sufficiently small and the data grows as $\mathcal{O}(\frac{1}{\epsilon^2})$, every solution $\hat{\Phi}$ to EIRM (equation 6) is in $\sqrt{\epsilon}$ radius of the OOD solution, i.e., $\|\hat{\Phi} - \tilde{S}^\mathsf{T}\gamma\| = \mathcal{O}(\sqrt{\epsilon})$. We contrast these rates to ones in the covariate shift setting (Section 3.3.1). Let $\mathbb{E}[Y^e|X^e = x] = \Psi^\mathsf{T} x$. If the data grows as $\mathcal{O}(\frac{1}{\nu^2})$, then both ERM and EIRM solution converge to Ψ as $\|\hat{\Phi} - \Psi\| = \mathcal{O}(\sqrt{\nu})$ (from Proposition 4). This shows that EIRM works in more settings (Proposition 4, 5) than ERM while matching the convergence rate of ERM.

2. Sample complexity grows polynomially in data dimension to ensure OOD generalization: Next, we set $\epsilon = \mu\epsilon_{\mathsf{th}}$ with $\mu \in [0, 1)$, and $|\mathcal{E}_{tr}| = 2n$ (satisfies Assumption 6 for $r = 1$). A simple manipulation of terms in Proposition 5 shows that sample complexity with quadratic growth in data dimension n, $\mathcal{O}\Big(\frac{n^2}{\mu^2}\log\big(\frac{2|\mathcal{H}_\Phi|}{\delta}\big)\Big)$, ensures $\hat{\Phi} = (\tilde{S}^\mathsf{T}\gamma)\alpha$ with $\alpha \in [\frac{1}{1+\sqrt{\mu}(\sqrt{2}-1)}, \frac{1}{1-\sqrt{\mu}(\sqrt{2}-1)}]$.

3. Comparison with Proposition 2: Lastly, we contrast sample complexity of EIRM in Proposition 5, $\mathcal{O}(\frac{1}{\epsilon^2})$, to Proposition 2, $\mathcal{O}(\max\{\frac{1}{\epsilon^2}, \frac{1}{\nu^2}\})$; the additional distributional assumptions in Proposition 5 help arrive at a lower sample complexity of $\mathcal{O}(\frac{1}{\epsilon^2})$. The bound in Proposition 2, $\mathcal{O}(\max\{\frac{1}{\epsilon^2}, \frac{1}{\nu^2}\})$, is larger than the one in Proposition 4, $\mathcal{O}(\frac{1}{\nu^2})$, and Proposition 5, $\mathcal{O}(\frac{1}{\epsilon^2})$, but is more general as it is agnostic to the distributional assumptions.

A simple illustration summarizing Propositions 4, 5: Set S to identity in Assumption 5. Recall Z_1^e and Z_2^e from Assumption 5. Since S is identity X^e can be written as $[X_1^e, X_2^e]$, where $X_1^e = Z_1^e$ and $X_2^e = Z_2^e$. If $X_2^e \perp \varepsilon^e$, then $\mathbb{E}[Y^e|X^e]$ is invariant and Assumption 4 holds. This corresponds to the setup in Figure 1a). We can now use Proposition 4 and deduce that ERM and IRM have same sample complexity and end up learning the ideal model that only uses the causal features X_1^e. If $X_2^e \leftarrow \varepsilon^e + N^e$, then this corresponds to the setup Figure 1b), X_1^e is the cause and X_2^e is spuriously correlated with label Y^e through the confounder ε^e. If $X_2^e \leftarrow Y^e + N^e$, then this corresponds to the setup in Figure 1c), X_1^e is the cause and X_2^e is anti-causally related to the label Y^e. In both these cases, the ideal OOD solution that solves equation 1 will only exploit X_1^e to make predictions. From Propositions 5, it follows that IRM when fed with $\mathcal{O}(\frac{1}{\epsilon^2})$ samples, it is in $\sqrt{\epsilon}$ radius of the target OOD solution, while ERM is asymptotically biased and exploits X_2^e (Proposition 17).

We define a polynomial version of the model in Assumption 5. We only need to change $Y^e \leftarrow \gamma^\mathsf{T}(Z_1^e) + \varepsilon^e$ to $Y^e \leftarrow \gamma^t \zeta_p(Z_1^e) + \varepsilon^e$. ζ_p is a polynomial feature map of degree p defined as $\zeta_p : \mathbb{R}^c \rightarrow \mathbb{R}^{c'}$, where c is the dimension of the input Z_1^e, $\zeta_p(W) = \big(W, W \otimes W, \ldots, (W \otimes W \ldots \text{p times} \otimes W)\big) = \big((W^{\otimes i})_{i=1}^p\big)$ and $\otimes$ is the Kronecker product. Also, $c' = \sum_{i=1}^p c^i$. Can we directly use the analysis from the linear case by transforming X^e appropriately? No, we first need to find an appropriate transformation for the scrambling matrix S that satisfies the conditions (invertibiltiy) in Assumption 5 while maintaining a linear relationship between transformations of X^e and Z^e. We present the main result informally below (details are in the supplement).

Proposition 6. *(Informal statement) For sufficiently small ϵ and $\delta \in (0, 1)$, if Assumptions similar to Proposition 5 hold and $|D| \geq \frac{16L'^4}{\epsilon^2} \log(\frac{2|\mathcal{H}_\Phi|}{\delta})$, then with a probability at least $1-\delta$, every solution $\hat{\Phi}$ to EIRM (equation 6) satisfies $\hat{\Phi} = \bar{\tilde{S}}^\mathsf{T}\gamma(\alpha)$, where $\bar{\tilde{S}}^\mathsf{T}\gamma$ is the OOD optimal model (defined in the supplement), $\alpha \in [\frac{1}{1+\tau\sqrt{\epsilon}}, \frac{1}{1-\tau\sqrt{\epsilon}}]$.*

Insights from the polynomial case and infinite hypothesis case. In the polynomial case, we adapt the linear general position Assumption 6, the number of environments $|\mathcal{E}_{tr}|$ are now required to grow as $\mathcal{O}(n^p)$. As a result, in the sample complexity analysis we discussed, we replace n with n^p to obtain that a sample complexity of $\mathcal{O}\Big(\frac{n^{2p}}{\mu^2} \log\big(\frac{2|\mathcal{H}_\Phi|}{\delta}\big)\Big)$ ensures $\hat{\Phi} = \bar{\tilde{S}}^\mathsf{T}\gamma(\alpha)$ with $\alpha \in [\frac{1}{1+\sqrt{\mu}(\sqrt{2}-1)}, \frac{1}{1-\sqrt{\mu}(\sqrt{2}-1)}]$. In the infinite hypothesis case, the main change in the results is that we replace $|\mathcal{H}_\Phi|$ with an appropriate model complexity metric (Shalev-Shwartz & Ben-David, 2014). Consider Proposition 5, a sample complexity of $\mathcal{O}\big(\frac{n^3}{\mu^2} \log\big(\frac{n}{\mu}\big)\big)$ ensures $\hat{\Phi} = (\tilde{S}^\mathsf{T}\gamma)\alpha$ with $\alpha \in [\frac{1}{1+\sqrt{\mu}(\sqrt{2}-1)}, \frac{1}{1-\sqrt{\mu}(\sqrt{2}-1)}]$ in contrast to $\mathcal{O}\Big(\frac{n^2}{\mu^2} \log\big(\frac{2|\mathcal{H}_\Phi|}{\delta}\big)\Big)$ in the finite hypothesis case. We showed the benefits of IRM for polynomial models and other extensions (non-linear S) are future work. In the supplement, we provide a dialogue explaining how our work fits in the big picture.

4 Experiments

In this section, we discuss classification experiments (regression experiments with similar qualitative findings are in the supplement). We introduce three new variants of the colored MNIST (CMNIST) dataset in (Arjovsky et al., 2019). We divide the training data in MNIST digits into two environments ($e = 1, 2$) equally and the testing data in MNIST digits is assigned to another environment ($e = 3$). X_g^e: gray scale image of the digit, Y_g^e: label of the gray scale digit (digits ≥ 5 have $Y_g^e = 1$ and digits < 5 have $Y_g^e = 0$). X^e: final colored image and Y^e: final label are generated as follows. Define Bernoulli variables G, N, N^e that take a value 1 with probability θ, β and β^e and 0 otherwise. Define a color variable C^e, where $C^e = 0$ is red and $C^e = 1$ is green. Let $\oplus$ denotes xor operation.

$$
\begin{aligned}
&Y_g^e \leftarrow \mathsf{L}(X_g^e), \mathsf{L} : \text{Human labeling,} \\
&Y^e \leftarrow \mathsf{L}(X_g^e) \oplus N, \text{ Corrupt the original labels with noise} \\
&C^e \leftarrow G(Y^e \oplus N^e) + (1-G)(N \oplus N^e), \text{ Use } G \text{ to select b/w anti-causal or confounded} \\
&X^e \leftarrow T(X_g, C^e), T : \text{transformation to color the image}
\end{aligned} \tag{10}
$$

If the probability $\theta = 1$, then $G = 1$ and that gives us back the original CMNIST in (Arjovsky et al., 2019), which we call anti-causal CMNIST (AC-CMNIST). If $\theta = 0$, then we get confounded colored MNIST (CF-CMNIST). If $0 < \theta < 1$, we get a hybrid dataset (HB-CMNIST). The above model in equation 10 has features of the model in Assumption 5, where X_g^e, C^e, L, T take the role of Z_1^e, Z_2^e, γ, S. We set the noise N parameter $\beta = 0.25$, and the parameter for N^e in the three environments $[\beta^1, \beta^2, \beta^3] = [0.1, 0.2, 0.9]$. Color is spuriously correlated with the label; $\mathbb{P}(C^e = 1|Y^e = 1)$ varies drastically in the three environments ($[0.9, 0.8, 0.1]$ for AC-CMNIST). In the variants of CMNIST we discussed, $\mathbb{P}(Y^e|X^e)$ varies across the environments. We now define a covariate shift based CMNIST (CS-CMNIST) ($\mathbb{P}(Y^e|X^e)$ is invariant). We use selection bias to induce spurious correlations. Generate a color C^e uniformly at random. Select the pair (X_g^e, C^e) with probability $1 - \psi^e$ if the label Y_g^e and C^e are the same, else select them with a probability ψ^e. If the pair is selected, then color the image $X^e \leftarrow T(X_g, C^e)$ and $Y^e \leftarrow Y_g^e$. Selection probability ψ^e for the three environments are $[\psi^1, \psi^2, \psi^3] = [0.1, 0.2, 0.9]$. Due to the selection bias, color is spuriously correlated, $\mathbb{P}(C^e = 1|Y^e = 1)$ varies drastically $[0.9, 0.8, 0.1]$. We provide the graphical models for the CMNIST variants and computations of $\mathbb{P}(C^e|Y^e)$, $\mathbb{P}(Y^e|X^e)$ in the supplement.

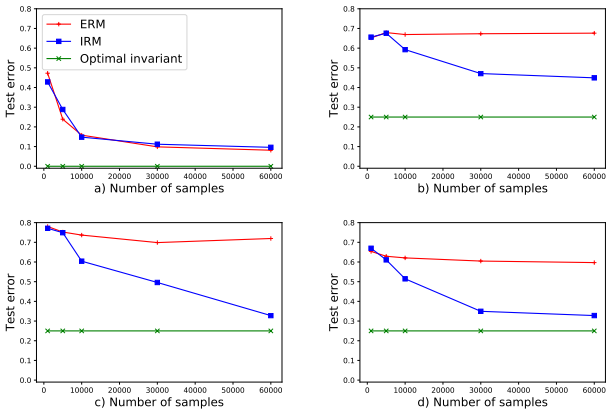

Figure 2: Comparisons: a) CS-CMNIST, b) CF-CMNIST, c) AC-CMNIST and d) HB-CMNIST.

4.1 Results

We use the first two environments ($e = 1, 2$) to train and third environment ($e = 3$) to test. Other details of the training (models, hyperparameters, etc.) are in the supplement. For each of the above datasets, we run the experiments for different amounts of training data from 1000 to up to 60000 samples (10 trials for each data size). In Figure 2, we compare the models trained using IRM and ERM in terms of the classification error on the test environment $e = 3$ (a poor performance indicates model exploits the color) for varying number of train samples. We also provide the performance of the ideal hypothetical optimal invariant model. Observe that except for in the covariate shift setting where IRM and ERM are similar as seen in Figure 2a (as predicted from Proposition 4), IRM outperforms ERM in the remaining three datasets (as predicted from Proposition 5) as seen in Figure 2b-d. We further validate this claim through the regression experiments provided in the supplement. In CF-CMNIST, IRM achievs an error of 0.45, which is much better than error of ERM (0.7) but is marginally better than a random guess. This suggests that confounder induced spurious correlations are harder to mitigate and may need more samples than needed in anti-causal case (AC-CMNIST).

5 Conclusion

We presented a sample complexity analysis of IRM to answer the question: when is IRM better than ERM (and vice-versa)? For distribution shifts such as the covariate shifts, we proved that both IRM and ERM have similar sample complexity and arrive at the desired OOD solution asymptotically. For distribution shifts involving confounders and (or) anti-causal variables and polynomial generative models, we proved that IRM is guaranteed to achieve the desired OOD solution while ERM can be asymptotically biased. We proposed new variants of original colored MNIST dataset from Arjovsky et al. (2019), which are more comprehensive and better capture how spurious correlations occur in reality. To the best of our knowledge, we believe this to be the first work that provides a rigorous characterization of impact of factors such as model complexity, number of environments on the sample complexity and its distance from the OOD solution in distribution shifts that go beyond covariate shifts.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization game. In *International Conference on Machine Learning*, 2020.

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

Isabela Albuquerque, João Monteiro, Tiago H Falk, and Ioannis Mitliagkas. Adversarial target-invariant representation learning for domain generalization. *arXiv preprint arXiv:1911.00804*, 2019.

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

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision*, pp. 456–473, 2018.

Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, pp. 139–153. Springer, 2012.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pp. 137–144, 2007.

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(1-2):151–175, 2010.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, 2020.

Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 129–136, 2010.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *medRxiv*, 2020.

Zhun Deng, Frances Ding, Cynthia Dwork, Rachel Hong, Giovanni Parmigiani, Prasad Patil, and Pragya Sur. Representation via representations: Domain generalization via adversarially learned invariant representations. *arXiv preprint arXiv:2006.11478*, 2020.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. *arXiv preprint arXiv:2008.01883*, 2020.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

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*, pp. 5542–5550, 2017.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. *arXiv preprint arXiv:2006.07500*, 2020.

Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, pp. 11749–11756, 2020.

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

Artidoro Pagnoni, Stefan Gramatovici, and Samuel Liu. Pac learning guarantees under covariate shift. *arXiv preprint arXiv:1812.06393*, 2018.

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

Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *International Conference on Machine Learning*, 2020.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. *Advances in Domain Adaptation Theory*. Elsevier, 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894*, 2020.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon. On learning invariant representation for domain adaptation. *arXiv preprint arXiv:1901.09453*, 2019.