Does Invariant Risk Minimization Capture Invariance?

Pritish Kamath pritish@ttic.edu
Akilesh Tangella akilesh@ttic.edu
Danica J. Sutherland dsuth@cs.ubc.ca
Nathan Srebro nati@ttic.edu
Toyota Technological Institute at Chicago

Abstract

We show that the Invariant Risk Minimization (IRM) formulation of Arjovsky et al. (2019) can fail to capture "natural" invariances, at least when used in its practical "linear" form, and even on very simple problems which directly follow the motivating examples for IRM. This can lead to worse generalization on new environments, even when compared to unconstrained ERM. The issue stems from a significant gap between the linear variant (as in their concrete method IRMv1) and the full non-linear IRM formulation. Additionally, even when capturing the "right" invariances, we show that it is possible for IRM to learn a sub-optimal predictor, due to the loss function not being invariant across environments. The issues arise even when measuring invariance on the population distributions, but are exacerbated by the fact that IRM is extremely fragile to sampling.

1 INTRODUCTION

Machine learning systems tend to seize on spurious correlations present in the training data, and so when presented with out-of-distribution inputs, they can fail spectacularly. For instance, in the spirit of Beery et al. (2018) and Arjovsky et al. (2019), consider a deep neural network trained to classify images as containing a cow or a camel. Suppose that most pictures of cows in the training set are taken in (green) grassy pastures, and those of camels are mostly in (brown) deserts. Then, the neural network is likely to strongly use background color for its predictions – after all, it is a very easy signal to use, and it barely hurts the loss. Such a network, however, will perform poorly at recognizing cows on a beach. How, then, can we design a machine learning system to identify key features of interest – face, shape, body color, etc., of animals – and ignore spurious ones, like the background color?

Standard machine learning algorithms assume a training set independently sampled from a *single* distribution, and seek good performance only on new samples from the same distribution. There has been much work on models that can adapt to a new distribution given a small number of labeled samples (see e.g. the survey of Redko et al. 2020), or models that are robust to *nearby* distributions (see e.g. the survey of Rahimian and Mehrotra 2019). Ideally though, we would hope for a model that can handle even *large* changes in distribution, *without* the need for labeled target samples.

In reality, our training data usually does *not* actually come from a single homogeneous source: we may have collected it from different users, on different continents, in different years. We thus may be able to tell which correlations are stable across environments (and hence are more likely to be the "true" correlations we seek), and which behave differently in different environments (and are more likely to be spurious).

One approach, then, is to attempt to learn an *invariant predictor* (e.g. Peters et al. 2015; Heinze-Deml et al. 2018; Rojas-Carulla et al. 2018). We might, for instance, assume that for the *causally relevant* subset S of the input variables X, the conditional distribution $\{Y|X_S\}$ is invariant across data sampled from different environments. This usually requires assuming a meaningful causal graph relating the observed variables. When classifying cows vs. camels based on image pixels, such assumptions are not likely to hold on the input data, though they could potentially apply to the latent variables underlying these images.

The *Invariant Risk Minimization* (IRM) framework of Arjovsky et al. (2019) tries to find a data representation φ which discards the spurious correlations, leaving only the "real" signal, by enforcing that the predictor w acting on that representation is simultaneously

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

optimal in each environment given φ. For instance, in the cows-vs-camels problem, φ might remove the background color. Since this gives a challenging bi-level optimization problem, Arjovsky et al. propose a relaxed version, IRMv1, which assumes w is a linear predictor. (We will overview the framework in Section 2.) For a thorough overview of how this approach fits into the literature on out-of-domain generalization, see the discussion by Arjovsky et al. and in particular Appendix A of Gulrajani and Lopez-Paz (2021). Subsequent work has provided new approaches for training in the IRM paradigm (e.g. Ahuja, Shanmugam, et al. 2020; Teney et al. 2020) and applications in domains such as interpretable language processing models (Chang et al. 2020).

Despite much initial promise, however, many key questions remain about the IRM framework: how well does IRMv1 approximate the exact version of the framework in general settings? Do invariant predictors always generalize well on unseen environments? When does a set of training environments allow us to find representations invariant across a broader set of target environments? How does the framework and/or the algorithm behave on finite samples?

Our Contributions We advance the understanding of several core questions about the IRM framework.

In Section 3, we study a simple setting of environments over $\mathcal{X} = \{0,1\}^2$, abstracting the `Colored-MNIST` problem studied by Arjovsky et al. (2019). We show that sometimes IRM with linear w can provably fail to find a "truly" invariant predictor, even when solved with respect to the population loss, and even if we provide *infinitely* many training environments. In fact, it finds a predictor that is even *worse* on out-of-distribution environments than unrestricted ERM. This issue persists in the IRMv1 implementation.

In Section 4, we note the population loss of even "truly" invariant predictors need not be invariant. We give a simple setting where IRM, which minimizes loss over training environments, prefers an invariant predictor with worse out-of-distribution generalization.

In Section 5, we study when it is possible to identify invariant predictors for a broad class of environments on the basis of a small range of training environments. Although this is generally impossible, we show conditions on the environments under which it is possible.

Finally, in Section 6, we point out issues that arise when using the IRM paradigm over the distributions of empirical samples rather than the population distributions. Here, even invariant predictors (over the population distributions) might not be invariant when considered over the distribution of empirical samples.

2 INVARIANT RISK MINIMIZATION

We now describe the IRM paradigm of Arjovsky et al. (2019). We have a set of *environments* $\mathcal{E}$, where each environment $e \in \mathcal{E}$ corresponds to a distribution $\mathcal{D}_e$ over $\mathcal{X} \times \mathcal{Y}$, with $\mathcal{X}$ being the space of inputs and $\mathcal{Y}$ that of outputs. Our goal is to find a predictor $f : \mathcal{X} \to \widehat{\mathcal{Y}}$; we measure the quality of a prediction with a loss function $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, and the quality of a predictor by its *population loss* on environment $e \in \mathcal{E}$, given by $\mathcal{L}_e(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}_e} \ell(f(x), y)$. In this paper, we mainly focus on the following special case.

Setting A. *$\mathcal{Y} \subseteq \mathbb{R}$, $\widehat{\mathcal{Y}} = \mathbb{R}$, and ℓ is either the square loss $\ell_{\text{sq}}(\widehat{y}, y) := \frac{1}{2}(\widehat{y} - y)^2$, or, when $\mathcal{Y} = \{-1, 1\}$ (corresponding to binary classification), the logistic loss $\ell_{\log}(\widehat{y}, y) := \log(1 + \exp(-\widehat{y}y))$.*

Given access to samples from some training environments $\mathcal{E}_{\text{tr}} \subseteq \mathcal{E}$, our aim to learn a predictor f that minimizes the "out-of-distribution" loss over all environments in $\mathcal{E}$, namely

$$\mathcal{L}_{\mathcal{E}}(f) := \sup_{e\in\mathcal{E}} \mathcal{L}_e(f). \qquad \text{(OOD-Gen)}$$

2.1 Notions of Invariance

The IRM paradigm attempts to solve this problem by learning an *invariant* representation $\varphi : \mathcal{X} \to \mathcal{Z}$. For instance, φ might "throw away" the spurious background color in the cows-vs.-camels example, if $e_1 \in \mathcal{E}$ is images from Ireland (where most cow images have grassy backgrounds), and $e_2 \in \mathcal{E}$ is from India (with many more images of cows on city streets). The formal definition of *invariant* is as follows.

Definition 1 (Definition 3 of Arjovsky et al. 2019)**.** *A representation[1] $\varphi : \mathcal{X} \to \mathcal{Z}$ is* invariant *over a set of environments $\mathcal{E}$ if there exists a $w : \mathcal{Z} \to \widehat{\mathcal{Y}}$ such that w is simultaneously optimal on φ for all environments $e \in \mathcal{E}$, that is, $w \in \operatorname{argmin}_{\overline{w}:\mathcal{Z}\to\widehat{\mathcal{Y}}} \mathcal{L}_e(\overline{w} \circ \varphi)$.*

This definition is motivated by the following observation of Arjovsky et al. (2019), which corresponds more closely to an intuitive definition of invariance.

Observation 2. *Under Setting A, a representation $\varphi : \mathcal{X} \to \mathcal{Z}$ is invariant over $\mathcal{E}$ if and only if for all $e_1, e_2 \in \mathcal{E}$, it holds that*

$$\mathbb{E}_{\mathcal{D}_{e_1}}[Y \mid \varphi(X) = z] = \mathbb{E}_{\mathcal{D}_{e_2}}[Y \mid \varphi(X) = z]$$

for all $z \in \mathcal{Z}_\varphi^{e_1} \cap \mathcal{Z}_\varphi^{e_2}$, where $\mathcal{Z}_\varphi^e$ are the representations from $\mathcal{D}_e$, $\mathcal{Z}_\varphi^e := \{\varphi(X) \mid (X, Y) \in \operatorname{Supp}(\mathcal{D}_e)\}$.

[1] We always assume φ and w are measurable. For further subtleties with Definitions 1 and 3, see Appendix A.1.

We give a proof in Appendix A.2 for completeness.

Crucially, Definition 1 requires that φ and w are unrestricted in the space of *all* (measurable) functions. However, we wish to learn φ and w with access to only (finite) training sets S_e sampled from $\mathcal{D}_e$, for only a small subset of training environments $\mathcal{E}_{\text{tr}} \subseteq \mathcal{E}$. For this to be feasible, it is natural to add a restriction that $\varphi \in \Phi$ and $w \in \mathcal{W}$, for suitable classes Φ of functions mapping $\mathcal{X} \to \mathcal{Z}$ and $\mathcal{W}$ of functions mapping $\mathcal{Z} \to \widehat{\mathcal{Y}}$. Any choice of function classes $(\Phi, \mathcal{W})$ defines a class of "invariant" predictors for a set of environments $\mathcal{E}$.

Definition 3. *For any Φ, $\mathcal{W}$ and loss function ℓ, the set of invariant predictors on $\mathcal{E}$, $\mathcal{I}^{\ell}_{\Phi,\mathcal{W}}(\mathcal{E})$, is the set of all predictors $f : \mathcal{X} \to \widehat{\mathcal{Y}}$ such that $\exists\, (w, \varphi) \in \mathcal{W} \times \Phi$ satisfying the following:*

- *$f = w \circ \varphi$, and*
- *for all $e \in \mathcal{E}$, $w \in \operatorname{argmin}_{\overline{w} \in \mathcal{W}} \mathcal{L}_e(\overline{w} \circ \varphi)$.*

For ease of notation, we will keep the loss function ℓ implicit. When Φ is the space of all functions $\mathcal{X} \to \mathcal{Z}$, we denote $\mathcal{I}_{\Phi,\mathcal{W}}(\mathcal{E})$ as simply $\mathcal{I}_{\mathcal{W}}(\mathcal{E})$. Moreover, when $\mathcal{W}$ is the space of all functions $\mathcal{Z} \to \widehat{\mathcal{Y}}$, we denote $\mathcal{I}_{\mathcal{W}}(\mathcal{E})$ as $\mathcal{I}(\mathcal{E})$, leaving the choice of $\mathcal{Z}$ implicit.[2]

Because exact optimization over $\mathcal{W}$ is in general difficult, it is useful to consider some special cases. A natural option is *linear* invariant predictors, where $\mathcal{Z} = \mathbb{R}^d$ and $\mathcal{W} = \mathcal{W}^d_{\text{lin}}$ is the space of all linear functions on $\mathbb{R}^d$. Arjovsky et al. (2019) argued that linear predictors in fact provide no additional representation advantage over *scalar* invariant predictors, the linear predictors for $d = 1$, $\mathcal{W} = \mathcal{S} := \mathcal{W}^1_{\text{lin}}$. In our notation, this translates to the following lemma, proved in Appendix A.2.

Lemma 4. *Under Setting A, for all $\mathcal{E}$ and $d \geq 1$,*

$$\mathcal{I}(\mathcal{E}) \;\subseteq\; \mathcal{I}_{\mathcal{S}}(\mathcal{E}) \;=\; \mathcal{I}_{\mathcal{W}^d_{\text{lin}}}(\mathcal{E}).$$

2.2 Algorithms

Armed with a notion of invariance, we still need a way to pick an invariant predictor based on training environments $\mathcal{E}_{\text{tr}} \subseteq \mathcal{E}$. Arjovsky et al. (2019) proposed the *Invariant Risk Minimization* objective given by

$$\min_{\substack{\varphi:\mathcal{X}\to\mathcal{Z} \\ w:\mathcal{Z}\to\widehat{\mathcal{Y}}}} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(w \circ \varphi)$$
$$\text{s.t.}\, \forall e \in \mathcal{E}_{\text{tr}},\; w \in \operatorname{argmin}_{\overline{w}:\mathcal{Z}\to\widehat{\mathcal{Y}}} \mathcal{L}_e(\overline{w} \circ \varphi),$$

which in our notation is equivalent to

$$\min_{f\in\mathcal{I}(\mathcal{E}_{\text{tr}})} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(f). \tag{IRM}$$

[2] In defining $\mathcal{I}(\mathcal{E})$, the choice of $\mathcal{Z}$ does not matter, as long as $\mathcal{Z}$ is large enough compared to $\mathcal{X}$; for instance, $\mathcal{Z} = \mathcal{X}$ is always a valid choice.

We can analogously define $\mathsf{IRM}_{\mathcal{W}}$ to choose a predictor $f \in \mathcal{I}_{\mathcal{W}}(\mathcal{E}_{\text{tr}})$, and $\mathsf{IRM}_{\Phi,\mathcal{W}}$ from $f \in \mathcal{I}_{\Phi,\mathcal{W}}(\mathcal{E}_{\text{tr}})$.

Characterizing $\mathcal{I}_{\mathcal{W}}(\mathcal{E}_{\text{tr}})$ is difficult in general; fortunately $\mathcal{I}_{\mathcal{W}^d_{\text{lin}}}(\mathcal{E}) = \mathcal{I}_{\mathcal{S}}(\mathcal{E})$ affords a simple characterization. Any predictor $f \in \mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}})$ can be written as $f(x) = w_* \varphi_*(x)$ for a scalar w_*. Without loss of generality, we can simply absorb the scalar w_* into $\varphi := w_* \varphi_*$, so that $f = 1 \cdot \varphi$. In Setting A, where the loss function is convex and differentiable, $f = 1 \cdot \varphi \in \mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}}) = \mathcal{I}_{\mathcal{W}^d_{\text{lin}}}(\mathcal{E}_{\text{tr}})$ if and only if

$$\text{for all } e \in \mathcal{E}_{\text{tr}}, \quad \nabla_{w|w=1} \mathcal{L}_e(w \cdot \varphi) = 0. \tag{∇_w}$$

Yet, $\mathsf{IRM}_{\mathcal{S}}$ remains a bi-level optimization problem. For practical purposes, Arjovsky et al. (2019) proposed to soften this hard constraint, giving the algorithm IRMv1 to approximate $\mathsf{IRM}_{\mathcal{S}}$:

$$\min_{\varphi:\mathcal{X}\to\mathbb{R}} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(\varphi) + \lambda \left|\nabla_{w|w=1} \mathcal{L}_e(w \cdot \varphi)\right|^2. \tag{IRMv1}$$

A natural baseline is the ERM algorithm, which simply minimizes the loss over training environments:

$$\min_{f:\mathcal{X}\to\widehat{\mathcal{Y}}} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(f). \tag{ERM}$$

While we referred to IRM, IRMv1 and ERM as "algorithms" above, there still remain two key details that make these impractical as stated: (i) the loss minimized refers to the *population loss*, to which we do not have direct access, and (ii) we are assuming that φ is unrestricted in the space of all functions. Arjovsky et al. (2019) attempt to remedy these issues in IRMv1 by (i) replacing the population loss by the corresponding empirical loss measured over training sets, and (ii) by optimizing φ over a sufficiently expressive parameterized model, such as a deep neural network, using gradient-based local search methods.

Nevertheless, as we discuss shortly, $\mathsf{IRM}_{\mathcal{S}}$ does not capture IRM even when operating on the population loss with unrestricted φ. Unless otherwise stated, we always consider IRM, $\mathsf{IRM}_{\mathcal{S}}$, IRMv1 and ERM as operating over population losses.

2.3 Related Work

Rosenfeld et al. (2021) demonstrate an example where there exists a *near-optimal* solution to the IRMv1 objective, that nearly matches performance of IRM on training environments, but does no better than ERM on environments that are "far" away from the training distributions. This example relies on environments which barely overlap, allowing the representation to simply "memorize" the training environments. Indeed, Ahuja, Wang, et al. (2021) argue that IRM can have

an advantage over ERM only when the support of the different environment distributions have a significant overlap. Gulrajani and Lopez-Paz (2021) find empirically that with current models and data augmentation techniques, ERM achieves state-of-the-art practical performance in domain generalization. Nagarajan et al. (2021), meanwhile, theoretically study the behavior of ERM for domain generalization.

Note that in prior work, $\text{IRM}_\mathcal{S}$/IRMv1 and IRM are often referred to interchangeably. As we demonstrate, $\text{IRM}_\mathcal{S}$ can behave very differently from IRM, even on simple examples that motivated the IRM approach.

3 COLORED-MNIST AND TWO-BIT ENVIRONMENTS

To illustrate the utility of the IRM approach and IRMv1 in particular, Arjovsky et al. (2019) introduced the `Colored-MNIST` problem, a synthetic task derived from `MNIST` (LeCun et al. 2010). While `MNIST` images are grayscale, in `Colored-MNIST` each image is colored either red or green in a way that correlates strongly (but spuriously) with the class label. Here ERM learns to exploit the color, and fails at test time when the direction of correlation with the color is reversed.

To understand the behavior of $\text{IRM}_\mathcal{S}$ and IRMv1 on `Colored-MNIST`, we study an abstract version based on two bits of input, where Y is the binary label to be predicted, X_1 corresponds to the label of the handwritten digit (0-4 or 5-9), and X_2 corresponds to the color (red or green). We represent each environment e with two parameters $\alpha_e, \beta_e \in [0, 1]$. The distribution $\mathcal{D}_e$ is defined as

$$\begin{aligned} Y &\leftarrow \text{Rad}(0.5), \\ X_1 &\leftarrow Y \cdot \text{Rad}(\alpha_e), \\ X_2 &\leftarrow Y \cdot \text{Rad}(\beta_e), \end{aligned} \qquad \text{(Two-Bit-Envs)}$$

where $\text{Rad}(\delta)$ is a random variable taking value -1 with probability δ and $+1$ with probability $1 - \delta$. For convenience, we denote an environment e as (α_e, β_e).

Following the experiments with `Colored-MNIST` as done by Arjovsky et al. (2019), we consider a set of environments $\mathcal{E}_\alpha := \{(\alpha, \beta_e) : 0 < \beta_e < 1\}$. It can be shown that there only two predictors in $\mathcal{I}(\mathcal{E}_\alpha)$, one being the trivial 0-predictor, and another that depends only on X_1 (see proof of Proposition 5 for details).

Motivating example of Arjovsky et al. (2019) Consider $\mathcal{E} = \mathcal{E}_{0.25}$ and $\mathcal{E}_{\text{tr}} = \{(0.25, 0.1), (0.25, 0.2)\}$. Focusing on the case of ℓ_{sq}, (ERM) on $\mathcal{E}_{\text{tr}}$ learns the predictor f_{ERM} that is (approximately) given by

f_{ERM}	$X_2 = 1$	$X_2 = -1$
$X_1 = 1$	0.8889	−0.3077
$X_1 = -1$	0.3077	−0.8889

;

the prediction clearly depends on X_2 as well as X_1. On each environment in $\mathcal{E}_{\text{tr}}$, the signal from X_2 is stronger than that from X_1, and so the binary predictor here can be summarized as $\text{sign}(f_{\text{ERM}}(X)) = \text{sign}(X_2)$. On the other hand, (IRM) chooses the predictor f_{IRM}

f_{IRM}	$X_2 = 1$	$X_2 = -1$
$X_1 = 1$	0.5	0.5
$X_1 = -1$	−0.5	−0.5

,

whose binary behavior is $\text{sign}(f_{\text{IRM}}(X)) = \text{sign}(X_1)$.

On $e \in \mathcal{E}_{\text{tr}}$, f_{ERM} achieves a lower loss than f_{IRM}, since it is using the more powerful signal X_2. But, if we evaluate the ability of these predictors to generalize far out of distribution to a case where the (spurious) correlation of X_2 has flipped entirely, $e = (0.25, 0.9)$, f_{ERM} will give the wrong (binary) prediction 90% of the time, and get square loss $\mathcal{L}_e(f_{\text{ERM}}) = 0.985$. This is far worse than f_{IRM}, which at $\mathcal{L}_e(f_{\text{IRM}}) = 0.375$ has not suffered at all compared to $\mathcal{E}_{\text{tr}}$. It is even worse than the trivial 0-predictor, $\mathcal{L}_e(f_0) = 0.5$.

It turns out that $\text{IRM}_\mathcal{S}$ also learns the predictor f_{IRM} here, demonstrating the utility of this relaxation of IRM. This raises a natural question:

Does $\text{IRM}_\mathcal{S}$ *always learn the same predictor as* IRM*?*

Arjovsky et al. (2019, Section 4.1) considered a specialized *linear* family of environments, where they proved that indeed $\text{IRM}_\mathcal{S}$ learns an invariant predictor, as learned by IRM, for any $\mathcal{E}_{\text{tr}}$ with a sufficient number of environments in "general position."[3] (See also Rosenfeld et al. 2021, Section 5.) It was left to future work whether $\text{IRM}_\mathcal{S}$ learns invariant predictors in the sense of IRM more generally as well.

A failure mode of $\text{IRM}_\mathcal{S}$ and IRMv1 We show that in fact for a simple set of two-bit environments, $\text{IRM}_\mathcal{S}$ finds a predictor worse than that learned by IRM, and even worse than the one learned by ERM.

This occurs, e.g., for $\mathcal{E} = \mathcal{E}_{0.1}$ with training environments $\mathcal{E}_{\text{tr}} = \{e_1 = (0.1, 0.2), e_2 = (0.1, 0.25)\}$. The learned predictors are (approximately) as follows.

f_{ERM}	$X_2 = 1$	$X_2 = -1$
$X_1 = 1$	0.9375	0.4464
$X_1 = -1$	−0.4464	−0.9375

[3] The problem (Two-Bit-Envs) does not fit the setting of their Theorem 9, because flipping signs cannot be phrased as independent additive noise.

f_{IRM}	$X_2 = 1$	$X_2 = -1$
$X_1 = 1$	0.8	0.8
$X_1 = -1$	−0.8	−0.8

$f_{\mathsf{IRM}_\mathcal{S}}$	$X_2 = 1$	$X_2 = -1$
$X_1 = 1$	0.9557	0.2943
$X_1 = -1$	−0.2943	−0.9557

X_1 is the stronger signal for Y in this $\mathcal{E}_{\text{tr}}$, and all of these predictors make the same binary predictions, but with differing amounts of confidence. Extrapolating to the same kind of test environment where the correlation of X_2 has flipped, $e_{\text{test}} = (0.1, 0.9)$, we observe the following (approximate) losses:

	f_{ERM}	f_{IRM}	$f_{\mathsf{IRM}_\mathcal{S}}$	f_0
$\mathcal{L}_{e_1}(\cdot)$	0.15	0.18	0.15	0.5
$\mathcal{L}_{e_2}(\cdot)$	0.16	0.18	0.17	0.5
$\mathcal{L}_{e_{\text{test}}}(\cdot)$	0.28	0.18	0.38	0.5

.

The relation between IRM and ERM is as expected: IRM trades slightly worse loss on the training environments for much better extrapolation to the distant environment $e_{\text{test}} = (0.1, 0.9) \in \mathcal{E}_{0.1}$. But while $\mathsf{IRM}_\mathcal{S}$ also suffers slightly on the training environments, it is even worse than ERM at extrapolation to e_{test}! The invariant feature X_1 is more correlated with Y than the non-invariant feature X_2 in all of the training environments, and yet $\mathsf{IRM}_\mathcal{S}$ depends on X_2 even *more* seriously than ERM does.

Moreover, this is not a carefully-selected pathological example that would go away with more training environments. In fact, $\mathsf{IRM}_\mathcal{S}$ chooses the same predictor even if we include *any number of* additional training environments $(0.1, \beta_e)$ for $\beta_e < 0.28$. Indeed, we show that for these two-bit environments $\mathcal{E}_\alpha$, any two training environments are sufficient to recover the set of all invariant predictors (proof in Appendix B).

Proposition 5. *Under Setting A, for all $\alpha \in (0, 1)$ and $\mathcal{E}_{\text{tr}} = \{e_1, e_2\}$ for any two distinct $e_1, e_2 \in \mathcal{E}_\alpha$,*

$$\textit{(i)}\ \mathcal{I}_\mathcal{S}(\mathcal{E}_{\text{tr}}) = \mathcal{I}_\mathcal{S}(\mathcal{E}_\alpha) \quad \textit{and} \quad \textit{(ii)}\ \mathcal{I}(\mathcal{E}_{\text{tr}}) = \mathcal{I}(\mathcal{E}_\alpha).$$

Thus, the issue is not just that we have don't have enough training environments. Rather, as we will now show, what $\mathsf{IRM}_\mathcal{S}$ determines to be an "invariant predictor" is broader than our intuitive sense – or IRM's notion – of what it means to be invariant.

Predictors in $\mathcal{I}_\mathcal{S}(\mathcal{E}_\alpha)$ Recall a predictor $f = 1 \cdot \varphi$ is in $\mathcal{I}_\mathcal{S}(\mathcal{E}_{\text{tr}})$ if and only if φ satisfies Equation (∇_w).

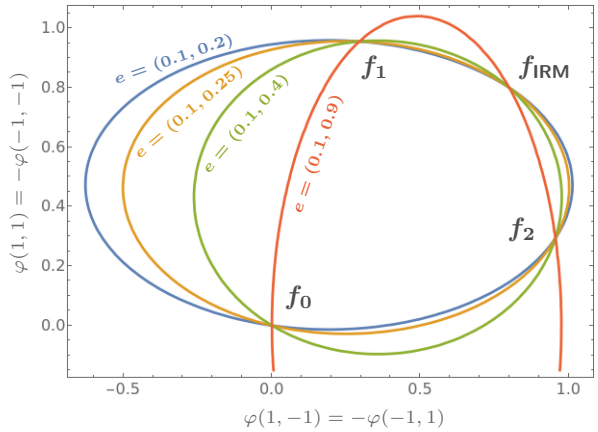


Figure 1: Odd solutions to (∇_w for ℓ_{sq}) for four environments in $\mathcal{E}_{0.1}$.

For ℓ_{sq}, this is same as having that for all $e \in \mathcal{E}_{\text{tr}}$,

$$\frac{\partial}{\partial w}\left(\mathbb{E}_{(X,Y)\sim\mathcal{D}_e}\ \frac{(w\cdot\varphi(X) - Y)^2}{2}\right)\Big|_{w=1} = 0\,,$$

or equivalently,

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}_e}\ (\varphi(X) - Y)\cdot\varphi(X) = 0. \quad (\nabla_w \text{ for } \ell_{\text{sq}})$$

This is a system of quadratic polynomials in four variables $\{\varphi(x) : x \in \{-1, 1\}^2\}$. For ease of visualization, we focus on *odd* predictors $f = 1 \cdot \varphi \in \mathcal{I}_\mathcal{S}(\mathcal{E}_{\text{tr}})$, namely those satisfying $f(x) = -f(-x)$ for all $x \in \{-1, 1\}^2$. This choice is motivated by the symmetry present in $\mathcal{D}_e$ and the loss ℓ_{sq}, along with the observation that the predictors f_{ERM}, f_{IRM} and $f_{\mathsf{IRM}_\mathcal{S}}$ are all odd. This allows us to focus on just two variables $\varphi(1, 1) = -\varphi(-1, -1)$ and $\varphi(1, -1) = -\varphi(-1, 1)$.

Figure 1 shows the solutions of (∇_w for ℓ_{sq}) among all odd φ for four environments in $\mathcal{E}_{0.1}$. There are precisely four odd choices of $\varphi \in \mathcal{I}_\mathcal{S}(\mathcal{E}_{0.1}) = \mathcal{I}_\mathcal{S}(\mathcal{E}_{\text{tr}})$. Two are the expected solutions f_0 and f_{IRM} described above; these are the only two predictors in $\mathcal{I}(\mathcal{E}_{\text{tr}}) = \mathcal{I}(\mathcal{E}_{0.1})$. $\mathcal{I}_\mathcal{S}(\mathcal{E}_{0.1})$, however, contains two *more* odd predictors, f_1 and f_2, the former being $f_{\mathsf{IRM}_\mathcal{S}}$ from above. $f_{\mathsf{IRM}_\mathcal{S}}$ achieves a smaller loss than the other solutions for the two training environments $(0.1, 0.2)$ and $(0.1, 0.25)$, but higher loss than f_{IRM} for environments $(0.1, 0.4)$ or $(0.1, 0.9)$. Figure 2 visualizes the losses of these four odd predictors on environments with varying β_e. Appendix B.1 has more details, including an analysis that explains precisely when these counterexamples arise.[4]

Thus, $\mathsf{IRM}_\mathcal{S}$ can find representations φ which are not *invariant* in the sense of Definition 1. In particular, for $\mathcal{E}_{0.1}$ with ℓ_{sq}, $\mathsf{IRM}_\mathcal{S}$'s feasible set of solutions is $\mathcal{I}_\mathcal{S}(\mathcal{E}_{\text{tr}}) \supsetneq \mathcal{I}(\mathcal{E}_{\text{tr}})$, or equivalently $\mathcal{I}_\mathcal{S}(\mathcal{E}_{0.1}) \supsetneq \mathcal{I}(\mathcal{E}_{0.1})$.

[4] This analysis was communicated to us by Léon Bottou.

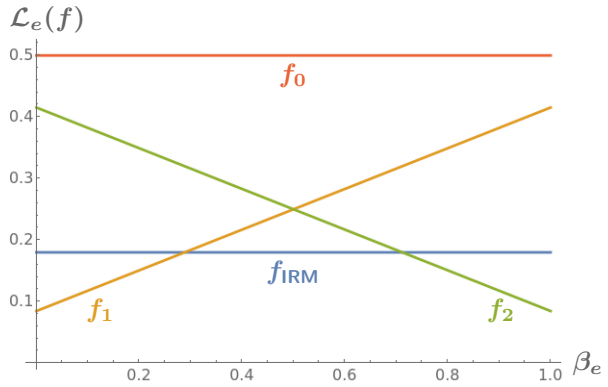


Figure 2: Losses $\mathcal{L}_e$ (for $\ell = \ell_{\text{sq}}$) of odd predictors in $\mathcal{I}_{\mathcal{S}}(\mathcal{E}_{0.1})$ for various $e = (0.1, \beta_e)$.

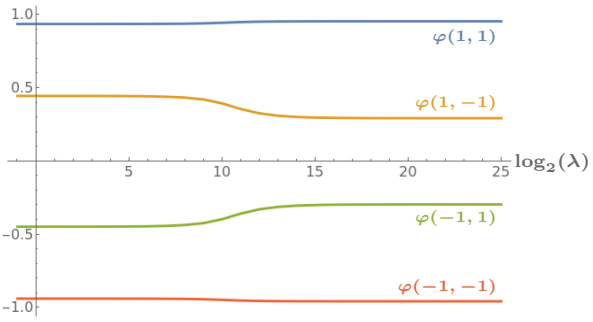


Figure 3: IRMv1 on $\mathcal{E}_{\text{tr}} = \{(0.1, 0.2), (0.1, 0.25)\}$. The horizontal axis is $\log_2(\lambda)$, with -1 representing $\lambda = 0$.

As seen from Figure 2, $f_{\text{IRM}_{\mathcal{S}}} = f_1$ has the lowest loss of those four solutions for $\beta_e \leq 0.28$. More training environments will not help $\text{IRM}_{\mathcal{S}}$ pick f_{IRM}, unless the average value of β_e across environments $e \in \mathcal{E}_{\text{tr}}$ is between 0.29 and 0.71. If the average value of β_e exceeds 0.72, $\text{IRM}_{\mathcal{S}}$ switches to the other solution f_2.

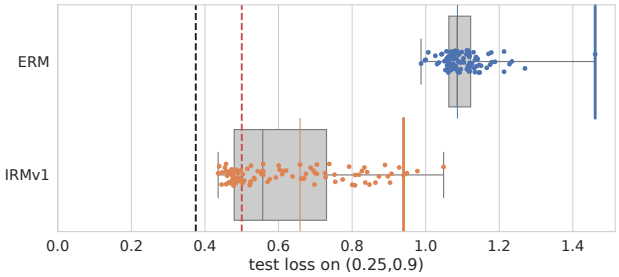
We know that IRMv1 becomes exactly ERM when its regularization weight is $\lambda = 0$, and $\text{IRM}_{\mathcal{S}}$ for $\lambda = \infty$. Figure 3 shows[5] the solution smoothly interpolating between f_{ERM} and $f_{\text{IRM}_{\mathcal{S}}}$, with the reliance on X_2 increasing as $\lambda \to \infty$.

ℓ_{log} loss A similar failure mode occurs for ℓ_{log} on $\mathcal{E}_{0.05}$ when training on $\mathcal{E}_{\text{tr}} = \{(0.05, 0.1), (0.05, 0.2)\}$. We give more details in Appendix B.2.

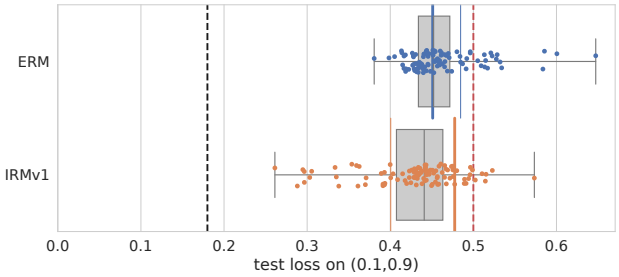
3.1 Experiments with `Colored-MNIST`

We now confirm that the failure mode studied above can also arise in practical training of deep networks based on IRMv1. `Colored-MNIST` corresponds to the two-bit environments above, where X_1 is a (grayscale) image from `MNIST`, and X_2 is a color (red or green) which is assigned to that image.[6] Thus, a learning algorithm which finds global minima of the IRMv1 population-level objective in a model capable of perfectly classifying `MNIST` digits would behave exactly as described above. In practice, however, we optimize empirical estimates of the risk and gradient penalty, in a model class which may not contain an exactly perfect digit classifier, with an algorithm which may not find the global optimum.

One significant practical issue with IRMv1 is in hyperparameter tuning, since we wish to find models which generalize to environments quite different from $\mathcal{E}_{\text{tr}}$. Arjovsky et al. (2019) chose hyperparameters arbitrarily for their ERM networks, and for IRMv1 by selecting a network with randomly selected hyperparameters which performed the best *on the test set* (specifically, the model with the highest minimum accuracy



(a) $\mathcal{E}_{\text{tr}} = \{(0.25, 0.1), (0.25, 0.2)\}$, $e_{\text{test}} = (0.25, 0.9)$



(b) $\mathcal{E}_{\text{tr}} = \{(0.1, 0.2), (0.1, 0.25)\}$, $e_{\text{test}} = (0.1, 0.9)$

Figure 4: Performance on e_{test} when training the given algorithm on $\mathcal{E}_{\text{tr}}$, using square loss ℓ_{sq}, with a fully-connected network. 100 repetitions are shown, using different random hyperparameters and training splits; boxplots show sample quartiles. Black dashed line (left) shows expected loss of the optimal invariant predictor f_{IRM}; red dashed line (right) shows expected loss of the other predictor in $\mathcal{I}(\mathcal{E}_{\text{tr}})$, the null predictor f_0. Shorter, colored vertical lines show the test set performance of the predictor which minimizes the training objective, (ERM) or (IRMv1) (with $\lambda = 10^6$).

[5] The IRMv1 objective can be non-convex, even for ℓ_{sq}, and typical optimization algorithms sometimes find local minima. We instead solved IRMv1 by explicitly enumerating the (odd) stationary points.

[6] In practice, we sample the image X_1 first and then flip Y with probability α_e; this is equivalent.

on $\mathcal{E}_{\text{tr}} \cup \{e_{\text{test}}\}$). Since this significantly advantages IRMv1 over ERM, we instead consider the distribution of performances with random hyperparameters from the same proposal distribution as used by Arjovsky et al. We also note which of these models minimized the objective on $\mathcal{E}_{\text{tr}}$ (using a fixed, large λ to compare the objective for IRMv1). Currently, there is no known principled approach for choosing λ; as noted by Gulrajani and Lopez-Paz (2021), this is often critical to the practical performance of IRM.

Arjovsky et al. (2019) use a fully-connected ReLU network with one hidden layer, operating on the red and green channels of a 14×14 image. Running ERM and IRMv1 on this architecture with ℓ_{sq} in the original `Colored-MNIST` problem shows (Figure 4a) that IRMv1 handily outperforms ERM in test loss, though it does not quite achieve the performance of the best possible f_{IRM}, and model selection based on $\mathcal{E}_{\text{tr}}$ would choose a predictor notably worse on the test set than the null predictor f_0. Moving to the example failure mode discussed above, this is no longer the case (Figure 4b): the two algorithms perform about the same in test loss, with model selection on $\mathcal{E}_{\text{tr}}$ selecting a model with performance about the same as f_0 for each algorithm. Although the practical instantiation of IRMv1 clearly suffers here, it is not *worse* than ERM as we would expect for the population-optimal solutions.

In this representation, X_1 (digit) and X_2 (color) are quite "entangled." In Appendix C, we consider an architecture which processes the grayscale image and total color of the image separately, thus becoming a little closer to the idealized setting (Two-Bit-Envs); here the failure of IRMv1 compared to ERM becomes more apparent. We also explore many variations of the experiment, including experiments with $\ell_{\log}$.

Thus, $\text{IRM}_{\mathcal{S}}$'s surprising failure on the extremely simple problem (Two-Bit-Envs) is essentially reproduced with practical optimization of neural networks on `Colored-MNIST`.

4 CAN IRM FAIL TO CHOOSE THE RIGHT PREDICTOR?

In the previous section, we saw an example where $\text{IRM}_{\mathcal{S}}$ was able to identify $\mathcal{I}_{\mathcal{S}}(\mathcal{E})$, since $\mathcal{I}_{\mathcal{S}}(\mathcal{E}) = \mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}})$ there, but chose a predictor in $\mathcal{I}_{\mathcal{S}}(\mathcal{E})$ with worse out-of-distribution risk for environments "far from" $\mathcal{E}_{\text{tr}}$. This happened because the loss $\mathcal{L}_e(f)$ of predictors $f \in \mathcal{I}_{\mathcal{S}}(\mathcal{E})$ need not be the same (invariant) for all environments $e \in \mathcal{E}$, and we pick the "wrong" predictor when optimizing $\sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$ over $f \in \mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}})$.

Is the same possible for IRM, or does its implicit premise that the optimal *invariant* predictor on $\mathcal{E}_{\text{tr}}$ will generalize well to $\mathcal{E}$ hold? IRM can of course fail when $\mathcal{I}(\mathcal{E}_{\text{tr}}) \supsetneq \mathcal{I}(\mathcal{E})$, when the training environments are not diverse enough to identify the right invariances. But what if we do have $\mathcal{I}(\mathcal{E}_{\text{tr}}) = \mathcal{I}(\mathcal{E})$?

The loss of an invariant predictor $f \in \mathcal{I}(\mathcal{E})$ need not be invariant for all $e \in \mathcal{E}$: consider e.g. varying amounts of inherent additive noise in a regression setting. This would still be acceptable as long as the *best* invariant predictor with respect to the population loss is the same for all environments $e \in \mathcal{E}$. Contrarily, we now give a simple family of environments $\mathcal{E}$, training environments $\mathcal{E}_{\text{tr}} \subseteq \mathcal{E}$ satisfying $\mathcal{I}(\mathcal{E}_{\text{tr}}) = \mathcal{I}(\mathcal{E})$, and two predictors $f_1, f_2 \in \mathcal{I}(\mathcal{E})$ such that $\mathcal{L}_e(f_1) > \mathcal{L}_e(f_2)$ for all $e \in \mathcal{E}_{\text{tr}}$, but $\mathcal{L}_{\mathcal{E}}(f_1) < \mathcal{L}_{\mathcal{E}}(f_2)$. Hence IRM prefers f_2 to f_1 based on $\mathcal{E}_{\text{tr}}$, but f_1 has better worst-case loss. It is thus generally difficult to handle out-of-distribution prediction in environments with more than one invariant predictor: the invariant predictor which is best on training environments might still perform poorly on unseen test environments, *despite being invariant*.

Consider environments $\mathcal{E}$ over $\mathcal{X} = \{-1, 0, 1\}^3$ and $\mathcal{Y} = \{-1, 1\}$, where each environment e is specified by a single parameter $\theta_e \in (-1/6, 1/3)$ as follows:

$$X_1 \leftarrow \begin{cases} -1 & \text{w.p. } \frac{1}{3} \\ 0 & \text{w.p. } \frac{1}{3} \\ +1 & \text{w.p. } \frac{1}{3} \end{cases}, \quad X_2 \leftarrow \begin{cases} -1 & \text{w.p. } \frac{1}{3} - \theta_e \\ 0 & \text{w.p. } \frac{1}{3} + 2\theta_e \\ +1 & \text{w.p. } \frac{1}{3} - \theta_e \end{cases},$$

$$\mathbb{E}_{\mathcal{D}_e}[Y | X_1, X_2] = 0.3(X_1 + X_2) + g_{\theta_e}(X_1, X_2) \ ,$$

where $g_{\theta_e}(x_1, x_2)$ is given as

$g_\theta(x_1, x_2)$	$x_2 = -1$	$x_2 = 0$	$x_2 = +1$
$x_1 = -1$	$\theta(\theta + \frac{2}{3})$	$-\theta(\frac{2}{3} - 2\theta)$	$3\theta^2$
$x_1 = 0$	$-\theta(\frac{2}{3} - 2\theta)$	0	$\theta(\frac{2}{3} - 2\theta)$
$x_1 = +1$	$-3\theta^2$	$\theta(\frac{2}{3} - 2\theta)$	$-\theta(\theta + \frac{2}{3})$

While the specific form of g_θ is a little involved, the main thing to note is that

$$\mathbb{E}_{\mathcal{D}_e}[g_{\theta_e}(X_1, X_2) \mid X_1] = 0 = \mathbb{E}_{\mathcal{D}_e}[g_{\theta_e}(X_1, X_2) \mid X_2]$$

which means that $\mathbb{E}_{\mathcal{D}_e}[Y|X_1] = 0.3X_1$ as well as $\mathbb{E}_{\mathcal{D}_e}[Y|X_2] = 0.3X_2$. Thus for ℓ_{sq}, $\mathcal{I}(\mathcal{E})$ contains the predictors $f_1(x) = 0.3\,x_1$ and $f_2(x) = 0.3\,x_2$. In fact, as shown in Appendix D, IRM will indeed pick among these predictors in $\mathcal{I}(\mathcal{E})$ for almost all $\mathcal{E}_{\text{tr}}$ containing at least two distinct environments:

Proposition 6. *In Setting A, for $\mathcal{E}$ as above, it holds for Lebesgue-almost all $\mathcal{E}_{\text{tr}} \subseteq \mathcal{E}$ with $|\mathcal{E}_{\text{tr}}| \geq 2$ that $\mathcal{I}(\mathcal{E}) = \mathcal{I}(\mathcal{E}_{\text{tr}})$. Moreover, any $f \in \mathcal{I}(\mathcal{E})$ depends on at most one of x_1 or x_2.*

Focusing on the case of ℓ_{sq}, the loss of the predictors

can be seen to be[7], for any $e \in \mathcal{E}$,

$$\mathcal{L}_e(f_1) = 0.47 \quad \text{and} \quad \mathcal{L}_e(f_2) = 0.47 + 0.09 \cdot \theta_e$$

Thus, if $\mathcal{E}_{\text{tr}}$ only contains environments e corresponding to $\theta_e < 0$, we will have that $\mathcal{L}_e(f_2) < \mathcal{L}_e(f_1)$ for all $e \in \mathcal{E}_{\text{tr}}$, and yet the invariant predictor that minimizes $\sup_{e \in \mathcal{E}} \mathcal{L}_e(\cdot)$ is f_1. See Figure 12 (Appendix D) for an illustration of these loss as a function of θ_e.

IRM's notion of invariance ensures $\mathbb{E}_{\mathcal{D}_e}[Y \mid \varphi(X) = z]$ is invariant across $\mathcal{E}$, but allows the loss of the corresponding predictor $\mathcal{L}_e(f)$ to differ across $e \in \mathcal{E}$. Here, in fact the full conditional distribution $\{Y \mid \varphi(X) = z\}$ is also invariant across $\mathcal{E}$, but even so, the loss varies. If we enforced a stronger notion of invariance which requires the entire joint distribution $\{(Y, \varphi(X))\}_{(X,Y)\sim\mathcal{D}_e}$ to be invariant across all $e \in \mathcal{E}$, we would not have faced this issue, since $\mathcal{L}_e$ would then be invariant, and indeed would pick f_1 in the example above. Yet this joint invariance is clearly too strict for some problems: it is impossible to achieve if the marginal distribution of Y differs across environments, and it is easy to construct other $\mathcal{E}$ where IRM allows the intuitively-correct predictor but joint invariance allows only a trivial constant predictor.

Thus, IRM is not always guaranteed to achieve optimal out-of-distribution loss, even when all the right invariances are captured by the training environments. The "right" notion of invariance really depends on what we know about the set of all environments $\mathcal{E}$.

5 WHEN DOES INVARIANCE GENERALIZE?

In the examples of Sections 3 and 4, it held that IRM or IRM$_{\mathcal{S}}$ were able to identify predictors invariant over all, even unseen, environments: specifically, $\mathcal{I}_{\mathcal{W}}(\mathcal{E}) = \mathcal{I}_{\mathcal{W}}(\mathcal{E}_{\text{tr}})$. That this holds is an implicit premise of the IRM framework. Yet it is unclear in general when invariances discovered on training environments will generalize to unseen environments. We now give some partial answers to this question.

For an arbitrary $\mathcal{E}$, we of course cannot expect invariances observed across $\mathcal{E}_{\text{tr}}$ to generalize over $\mathcal{E}$: simply consider adding a single entirely "irrelevant" e to $\mathcal{E}$. To provide some structure, we consider parameterized sets of environments $\mathcal{E}$. For simplicity, we focus on finite $\mathcal{X}$ and $\mathcal{Y}$, with $\mathcal{Y} \subseteq \mathbb{R}$. Let $\Delta_{\mathcal{X}\times\mathcal{Y}}$ denote the space of all probability distributions over $\mathcal{X} \times \mathcal{Y}$, and let $\Theta \subseteq \mathbb{R}^d$. A map $\Pi : \Theta \to \Delta_{\mathcal{X}\times\mathcal{Y}}$ naturally defines a set of environments $\mathcal{E}_\Pi$ corresponding to the set of distributions $\{\Pi(\theta) : \theta \in \Theta\}$. For example, the two-bit environments $\mathcal{E}_\alpha$ of Section 3 are parameterized by the map $\Pi : \theta \mapsto e = (\alpha, \theta)$, for $\theta \in \Theta = (0, 1)$.

$$\textit{For } \Theta_{\text{tr}} \subseteq \Theta \textit{ and } \mathcal{E}_{\text{tr}} = \{\Pi(\theta) \mid \theta \in \Theta_{\text{tr}}\},$$
$$\textit{when does it hold that } \mathcal{I}(\mathcal{E}_{\text{tr}}) = \mathcal{I}(\mathcal{E}_\Pi)?$$

Note that $\mathcal{I}(\mathcal{E}_\Pi) \subseteq \mathcal{I}(\mathcal{E}_{\text{tr}})$ always holds, but for any hope of $\mathcal{I}(\mathcal{E}_{\text{tr}}) \subseteq \mathcal{I}(\mathcal{E}_\Pi)$, we must assume $\mathcal{E}_{\text{tr}}$ contains a "representative set" of environments from $\mathcal{E}_\Pi$.

The most basic assumption to begin with is simply that Π is continuous. This is insufficient to guarantee invariance, even for very large Θ_{tr}: the map might simply "change directions" outside of $\mathcal{E}_{\text{tr}}$. We give a simple example below (proof in Appendix E), where even an uncountable number of environments in $\mathcal{E}_{\text{tr}}$ do not allow us to understand the full behavior of $\mathcal{E}$.

Proposition 7. *There exists a continuous map $\Pi : (0, 1) \to \Delta_{\mathcal{X}\times\mathcal{Y}}$ such that for $\Theta_{\text{tr}} = \left(0, \frac{1}{4}\right)$ and $\mathcal{E}_{\text{tr}} = \Pi(\Theta_{\text{tr}})$, it holds that $\mathcal{I}(\mathcal{E}_{\text{tr}}) \neq \mathcal{I}(\mathcal{E}_\Pi)$.*

On the other hand, if Π is not only continuous but also analytic, we *can* guarantee, under some conditions, that invariances over $\mathcal{E}_{\text{tr}}$ continue to hold over all of $\mathcal{E}_\Pi$. Let $\Pi_{(x,y)}(\theta) := \Pr_{(X,Y)\sim\Pi(\theta)}[X = x, Y = y]$ for each $(x, y) \in \mathcal{X}\times\mathcal{Y}$. We say the map $\Pi : \Theta \to \Delta_{\mathcal{X}\times\mathcal{Y}}$ is *analytic* if, for each $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\Pi_{(x,y)} : \Theta \to [0, 1]$ is analytic in θ.

Proposition 8. *Let $\Theta_{\text{tr}} \subseteq \Theta \subseteq \mathbb{R}^d$, where Θ is a connected, open set. Suppose $\Pi : \Theta \to \Delta_{\mathcal{X}\times\mathcal{Y}}$ is analytic, $\mathcal{X}$ and $\mathcal{Y}$ are finite and $\mathcal{E}_{\text{tr}} = \Pi(\Theta_{\text{tr}})$. Then, under Setting A,*

(i) For almost all Θ_{tr} with $|\Theta_{\text{tr}}| \geq 2$: $\mathcal{I}(\mathcal{E}_{\text{tr}}) = \mathcal{I}(\mathcal{E}_\Pi)$.

(ii) For all Θ_{tr} with non-zero Lebesgue measure: $\mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}}) = \mathcal{I}_{\mathcal{S}}(\mathcal{E}_\Pi)$.

The key step is that when Π is analytic, the conditional expectations $\mathbb{E}_{\Pi(\theta)}[Y \mid \varphi(X) = z]$ and the gradient $\nabla_{w|w=1}\mathcal{L}_{\Pi(\theta)}(w \cdot \varphi)$ are analytic functions in θ; the result is far stronger, however, for $\mathcal{I}$ (where the set of representations is finite) than for $\mathcal{I}_{\mathcal{S}}$, where our analysis requires uncountably many training environments. A version of Proposition 8 holds even for infinite spaces $\mathcal{X}$ and $\mathcal{Y} \subseteq \mathbb{R}$, under a technical definition of analyticity of Π (details in Appendix E), although in this case our result for $\mathcal{I}$ also requires $\mathcal{E}_{\text{tr}}$ to have positive measure.

Recall that the examples studied in Sections 3 and 4 indeed had analytic parameterizations, and hence Proposition 8 implies that $\mathcal{I}(\mathcal{E}_{\text{tr}}) = \mathcal{I}(\mathcal{E})$ holds for (almost) all $\mathcal{E}_{\text{tr}}$ with at least two distinct environments.

[7] This calculation does not need the specific form of g_θ.

6 IRM WITH FINITE SAMPLES

Except for Section 3.1, we have so far only discussed algorithms (IRM, $\mathsf{IRM}_\mathcal{S}$, and IRMv1) defined in terms of the *population* losses of training environments. In practice, however, we need to work with a finite number of samples from each training environment. If we directly apply IRM or $\mathsf{IRM}_\mathcal{S}$ as stated in (IRM) to empirical distributions, all correlations will have a small amount of noise, and it is extremely likely that the set of invariant predictors becomes empty.

On the other hand, IRMv1 for a fixed λ could be robust to sampling. We illustrate this in the two-bit environments of Section 3. Consider training environments $\mathcal{E}_{\text{tr}} = \{(0.25, 0.1), (0.25, 0.2), (0.25, 0.3)\}$: both IRM and $\mathsf{IRM}_\mathcal{S}$ are able to learn an invariant predictor. However, when sampling finite datasets, we only have that the empirical distribution of the two environments will be *close* to – but not exactly the same as – the true distribution; there may not be *any* exactly-invariant predictors. We illustrate this by evaluating $\mathsf{IRM}_\mathcal{S}$ on a set of training environments $\mathcal{E}'_{\text{tr}} = \{(0.245, 0.105), (0.255, 0.195), (0.251, 0.302)\}$, as a proxy for empirical distributions we see from finite samples. $\mathsf{IRM}_\mathcal{S}$ learns the trivial 0 predictor f_0; Figure 5 shows the behavior of IRMv1 for increasing λ.

For a fixed empirical distribution, it is likely that as $\lambda \to \infty$, IRMv1 approaches $\mathsf{IRM}_\mathcal{S}$, and does not find a good invariant predictor. If we instead take $n \to \infty$ for a fixed λ, though, we should approach the population version of IRMv1, and hence taking $\lambda \to \infty$ at an appropriate rate as $n \to \infty$ may approach the population $\mathsf{IRM}_\mathcal{S}$ predictor. Ahuja, Wang, et al. (2021) recently considered a variant of $\mathsf{IRM}_\mathcal{S}$ where the constraints (∇_w) defining $\mathcal{I}_\mathcal{S}(\mathcal{E}_{\text{tr}})$ need to hold ε-approximately. When training on the objective with finite samples, they bounds the sample complexity to get an out-of-distribution loss close to that of the corresponding population version of this ε-$\mathsf{IRM}_\mathcal{S}$.

Given the discrepancy between $\mathsf{IRM}_\mathcal{S}$ and IRM as pointed out in Section 3, however, it is important to make IRM itself more robust to finite samples. For instance, one possible approach would be to relax the requirement of $w \in \operatorname{argmin}_{\overline{w}:\mathcal{Z}\to\widehat{\mathcal{Y}}} \mathcal{L}_e(\overline{w} \circ \varphi)$ to

$$\mathcal{L}_e(w \circ \varphi) \;\le\; \min_{\overline{w}:\mathcal{Z}\to\widehat{\mathcal{Y}}} \mathcal{L}_e(\overline{w} \circ \varphi) + \varepsilon$$

for a suitable $\varepsilon > 0$. How to practically implement a version of this ε-IRM remains an open challenge.

7 DISCUSSION

The IRM framework of Arjovsky et al. (2019) proposes a promising new paradigm of learning, which attempts to exploit information we usually ignore to find models robust to even some quite dramatic changes in the input distribution. We have helped shed light on the applicability of this framework.

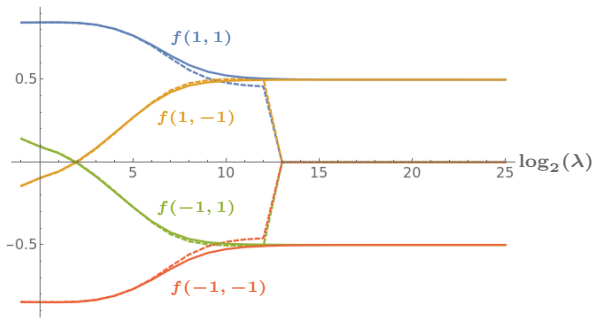


Figure 5: IRMv1 algorithm on exact environments $\mathcal{E}_{\text{tr}}$ (solid lines), and a noisy set $\mathcal{E}'_{\text{tr}}$ (dashed lines; definitions in text). The horizontal axis is $\log_2(\lambda)$, with -1 for $\lambda = 0$. Results are similar for small λ, until the noisy set abruptly gives the 0-predictor.

We now know that $\mathsf{IRM}_\mathcal{S}$ and IRMv1 can be surprisingly different from IRM, even on very simple environments. This emphasizes the importance of finding practical algorithms to approximate $\mathsf{IRM}_\mathcal{W}$ for some nonlinear class of functions $\mathcal{W}$.

We also know that even for IRM, choosing *among* invariant predictors can also be vital for out-of-domain generalization, and there exist cases where these algorithms choose the wrong one for out-of-distribution robustness. This holds even if we insist on a stronger notion of invariance, namely that of the conditional distribution $\{Y \mid \varphi(X)\}_{(X,Y)\sim\mathcal{D}_e}$. To truly handle worst-case out-of-distribution generalization, a stronger notion is needed: for example, it suffices to require invariance of the joint distribution $\{(Y, \varphi(X))\}_{(X,Y)\sim\mathcal{D}_e}$, but this seems overly stringent.

We also now know more about the possibility of generalizing invariances learned from $\mathcal{E}_{\text{tr}}$ to a larger set of environments $\mathcal{E}$. With significant structure on $\mathcal{E}$, it is possible to ensure $\mathcal{I}(\mathcal{E}) = \mathcal{I}(\mathcal{E}_{\text{tr}})$, but substantial questions remain as to the situation for $\mathcal{I}_\mathcal{S}$ or more realistic assumptions on $\mathcal{E}$.

Finally, we demonstrated that IRM and even IRMv1 can be surprisingly brittle when run on samples, rather than populations. Thus more analysis, and perhaps new algorithms, are needed to realize the promise of this framework in practice.

Acknowledgments

The authors would like to thank Léon Bottou, Martin Arjovksy, Ishaan Gulrajani, and David Lopez-Paz for useful discussions, particularly the derivation of the form of predictors in Appendix B.1.1.

Work was supported in part by NSF BIGDATA award 1546500 and NSF RI award 1764032. Work done while the authors participated in a special quarter on the Theory of Deep Learning sponsored by NSF TRIPOD award 1934843 (IDEAL) and while the first author participated in the Theory of Reinforcement Learning program at the Simons Institute for the Theory of Computing.

References

Ahuja, Kartik, Karthikeyan Shanmugam, Kush R. Varshney, and Amit Dhurandhar (2020). "Invariant Risk Minimization Games." *International Conference on Machine Learning*. arXiv: 2002.04692.

Ahuja, Kartik, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R. Varshney (2021). "Empirical or Invariant Risk Minimization? A Sample Complexity Perspective." *International Conference on Learning Representations*. arXiv: 2010.16412.

Arjovsky, Martin, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz (2019). *Invariant Risk Minimization*. arXiv: 1907.02893.

Beery, Sara, Grant Van Horn, and Pietro Perona (2018). "Recognition in Terra Incognita." *15th European Conference on Computer Vision*. DOI: 10.1007/978-3-030-01270-0_28.

Chang, Shiyu, Yang Zhang, Mo Yu, and Tommi S. Jaakkola (2020). "Invariant Rationalization." *International Conference on Machine Learning*. arXiv: 2003.09772.

Gulrajani, Ishaan and David Lopez-Paz (2021). "In Search of Lost Domain Generalization." *International Conference on Learning Representations*. arXiv: 2007.01434.

Heinze-Deml, Christina, Jonas Peters, and Nicolai Meinshausen (2018). "Invariant Causal Prediction for Nonlinear Models." *Journal of Causal Inference* 6.2, p. 20170016. DOI: 10.1515/jci-2017-0016.

LeCun, Yann, Corinna Cortes, and CJ Burges (2010). "MNIST handwritten digit database." *ATT Labs [Online]* 2. URL: http://yann.lecun.com/exdb/mnist.

Mityagin, Boris (2015). *The Zero Set of a Real Analytic Function*. arXiv: 1512.07276.

Nagarajan, Vaishnavh, Anders Andreassen, and Behnam Neyshabur (2021). "Understanding the failure modes of out-of-distribution generalization." *International Conference on Learning Representations*. arXiv: 2010.15775.

Peters, Jonas, Peter Bühlmann, and Nicolai Meinshausen (2015). "Causal inference using invariant prediction: identification and confidence intervals." *Journal of the Royal Statistical Society, Series B* 78.5, pp. 947–1012. DOI: 10.1111/rssb.12167. arXiv: 1501.01332.

Planet Math (Mar. 22, 2013). *Differentiation under the Integral Sign*. URL: https://planetmath.org/differentiationundertheintegralsign.

Rahimian, Hamed and Sanjay Mehrotra (2019). *Distributionally Robust Optimization: A Review*. arXiv: 1908.05659.

Redko, Ievgen, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani (2020). *A survey on domain adaptation theory: learning bounds and theoretical guarantees*. arXiv: 2004.11829.

Rojas-Carulla, Mateo, Bernhard Schölkopf, Richard Turner, and Jonas Peters (2018). "Invariant Models for Causal Transfer Learning." *Journal of Machine Learning Research* 19.36, pp. 1–34. arXiv: 1507.05333.

Rosenfeld, Elan, Pradeep Kumar Ravikumar, and Andrej Risteski (2021). "The Risks of Invariant Risk Minimization." *International Conference on Learning Representations*. arXiv: 2010.05761.

Teney, Damien, Ehsan Abbasnejad, and Anton van den Hengel (2020). *Unshuffling Data for Improved Generalization*. arXiv: 2002.11894.