THE RISKS OF INVARIANT RISK MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Invariant Causal Prediction (Peters et al., 2016) is a technique for out-of-distribution generalization which assumes that some aspects of the data distribution vary across the training set but that the underlying causal mechanisms remain constant. Recently, Arjovsky et al. (2019) proposed Invariant Risk Minimization (IRM), an objective based on this idea for learning deep, invariant features of data which are a complex function of latent variables; many alternatives have subsequently been suggested. However, formal guarantees for all of these works are severely lacking. In this paper, we present the first analysis of classification under the IRM objective—as well as these recently proposed alternatives—under a fairly natural and general model. In the linear case, we show simple conditions under which the optimal solution succeeds or, more often, fails to recover the optimal invariant predictor. We furthermore present the very first results in the non-linear regime: we demonstrate that IRM can fail catastrophically unless the test data are sufficiently similar to the training distribution—this is precisely the issue that it was intended to solve. Thus, in this setting we find that IRM and its alternatives fundamentally *do not improve* over standard Empirical Risk Minimization.

1 INTRODUCTION

Prediction algorithms are evaluated by their performance on unseen test data. In classical machine learning, it is common to assume that such data are drawn i.i.d. from the same distribution as the data set on which the learning algorithm was trained—in the real world, however, this is often not the case. When this discrepancy occurs, algorithms with strong in-distribution generalization guarantees, such as Empirical Risk Minimization (ERM), can fail catastrophically. In particular, while deep neural networks achieve superhuman performance on many tasks, there is evidence that they rely on statistically informative but non-causal features in the data (Beery et al., 2018; Geirhos et al., 2018; Ilyas et al., 2019). As a result, such models are prone to errors under surprisingly minor distribution shift (Su et al., 2019; Recht et al., 2019). To address this problem, researchers have investigated alternative objectives for training predictors which are robust to possibly egregious shifts in the test distribution.

The task of generalizing under such shifts, known as *Out-of-Distribution (OOD) Generalization*, has led to many separate threads of research. One approach is Bayesian deep learning, accounting for a classifier's uncertainty at test time (Neal, 2012). Another technique that has shown promise is data augmentation—this includes both automated data modifications which help prevent overfitting (Shorten & Khoshgoftaar, 2019) and specific counterfactual augmentations to ensure invariance in the resulting features (Volpi et al., 2018; Kaushik et al., 2020).

A strategy which has recently gained particular traction is Invariant Causal Prediction (ICP; Peters et al. 2016), which views the task of OOD generalization through the lens of causality. This framework assumes that the data are generated according to a Structural Equation Model (SEM; Bollen 2005), which consists of a set of so-called mechanisms or structural equations that specify variables given their parents. ICP assumes moreover that the data can be partitioned into *environments*, where each environment corresponds to interventions on the SEM (Pearl, 2009) on just the *non-causal mechanisms*, and without modifying those mechanisms that are causal. The justification for this is that the causal mechanisms of the data generating process are unchanging but other effects can vary. Therefore, learning relationships that are invariant across environments ensures recovery of the invariant features which generalize under arbitrary interventions. In this work, we consider objectives that attempt to learn what we refer to as the "invariant classifier"—this is the classifier which uses and is optimal with respect to only the invariant features in the SEM. By definition, such a classifier does not overfit to environment-specific properties of the data distribution, so it will generalize even under major distribution shift at test time. In particular, we focus our analysis on one of the more popular objectives, Invariant Risk Minimization (IRM; Arjovsky et al. (2019)), but our results can easily be extended to similar recently proposed alternatives (see Appendix E).

Various works on invariant prediction (Muandet et al., 2013; Ghassami et al., 2017; Heinze-Deml et al., 2018; Rojas-Carulla et al., 2018; Subbaswamy et al., 2019; Christiansen et al., 2020) consider regression in both the linear and

non-linear setting, but they exclusively focus on learning with partially observed covariates or instrumental variables. Under such a condition, results from causal inference (Maathuis et al., 2009; Peters et al., 2017) allow for formal guarantees of the identification of the invariant features, or at least a strict subset of them. With the rise of deep learning, more recent literature has developed objectives for learning invariant representations when the data are a non-linear function of unobserved latent factors, a common assumption when working with complex, high-dimensional data such as images. Causal discovery and inference with unobserved confounders or latents is a much harder problem (Peters et al., 2017), so while empirical results seem encouraging, these objectives are presented with few formal guarantees.

IRM is one such objective for invariant representation learning. The goal of IRM is to learn a feature embedder such that the optimal linear predictor on top of these features is the same for every environment—the idea being that only the invariant features will have an optimal predictor that is invariant. Thus, IRM hopes to "extrapolate" to new test environments, unlike ERM which excels at "interpolating". Recent works have pointed to shortcomings of IRM and have suggested modifications which they claim prevent these failures. However, these alternatives are compared in broad strokes, with little in the way of theory.

In this work, we present the first formal analysis of classification under the IRM objective under a fairly natural and general model which is similar to the original work. Our results show that despite being inspired by invariant prediction, this objective can frequently be expected to perform *no better than ERM*. In the linear setting, we present simple conditions under which solving to optimality succeeds or, more often, breaks down in recovering the invariant classifier. We additionally demonstrate a fundamental failure case—under mild conditions, there exists a feasible point that uses only non-invariant features and achieves lower empirical risk than the invariant classifier; thus it will appear as a more attractive solution, yet its reliance on non-invariant features mean it will completely fail to generalize. As corollaries, we present similar cases where Risk Extrapolation (REx; Krueger et al. 2020) and similar objectives likewise fail. Futhermore, we present the *first results in the non-linear regime*: we demonstrate the existence of a classifier with exponentially small sub-optimality which nevertheless heavily relies on non-invariant features on most test inputs, resulting in worse-than-chance performance on distributions that are sufficiently dissimilar from the training environments. These findings strongly suggest that existing approaches to ICP for high-dimensional latent variable models do not cleanly achieve their stated objective and that future work would benefit from a more formal treatment.

2 Related work

Works on learning deep invariant representations vary considerably in their model assumptions. Some works search for a *domain-invariant* representation (Muandet et al., 2013; Ganin et al., 2016), i.e. invariance of the covariate distribution $p(\Phi(x))$, but this is typically used for domain adaptation (Ben-David et al., 2010; Ganin & Lempitsky, 2015; Zhang et al., 2015; Long et al., 2018), with assumed access to labeled or unlabeled data from the target distribution. Further, there is evidence that this condition may not be sufficient for effective domain adaptation or OOD generalization (Zhao et al., 2019; Johansson et al., 2019). Other works instead hope to find representations that are *conditionally* domain-invariant, expecting invariance of $p(\Phi(x) \mid y)$ (Gong et al., 2016; Li et al., 2018).

More recent works, including the objectives discussed in this paper, suggest invariance of the *feature-conditioned label distribution*. In particular, Arjovsky et al. (2019) only assume invariance of $\mathbb{E}[y \mid \Phi(x)]$; follow-up works rely on a stronger assumption of invariance of $p(y \mid \Phi(x))$ (Krueger et al., 2020; Xie et al., 2020; Jin et al., 2020; Mahajan et al., 2020; Bellot & van der Schaar, 2020). Though this approach has become popular in the last year, it is somewhat similar to the existing concept of *covariate shift* (Shimodaira, 2000; Bickel et al., 2009), which considers the same setting. The main difference is that these more recent works assume that the shifts in $p(\Phi(x))$ occur between discrete, labeled environments, as opposed to more generally from train to test distributions.

3 Our Results

We consider a structural causal model with explicit separation of *invariant* features z_c, whose joint distribution with the label is fixed for all environments, and *environmental* features z_e ("non-invariant"), whose distribution can vary across different environments. We assume that data are drawn from a set of E training environments $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$ and that we know from which environment each sample is drawn. For a given environment e, the data are defined by the following process: first, a label $y \in \{\pm 1\}$ is drawn according to a fixed probability:

$$y = \begin{cases} 1, & \text{w.p. } \eta \\ -1, & \text{otherwise.} \end{cases} \tag{1}$$

Next, both invariant features and environmental features are drawn according to a Gaussian[1]:

$$z_c \sim \mathcal{N}(y \cdot \mu_c, \sigma_c^2 I), \qquad z_e \sim \mathcal{N}(y \cdot \mu_e, \sigma_e^2 I), \tag{2}$$

with $\mu_c \in \mathbb{R}^{d_c}, \mu_e \in \mathbb{R}^{d_e}$—typically, for complex, high-dimensional data we would expect $E < d_c \ll d_e$. Finally, the observation x is generated as a function of the latent features:

$$x = f(z_c, z_e). \tag{3}$$

We assume f is injective, so that it is in principle possible to recover the latent features from the observations, i.e. there exists a function Φ such that $\Phi(f(z_c, z_e)) = [z_c, z_e]^T$. We write the joint and marginal distributions as $p^e(x, y, z_c, z_e)$. Where it is clear from context, we omit the specific arguments.

Remarks on the model. This model is natural and flexible; it generalizes several existing models used to analyze learning under the existence of adversarial distribution shift or non-invariant correlations (Schmidt et al., 2018; Sagawa et al., 2020). The fundamental facet of this model is the constancy of the invariant parameters $\eta, \mu_c, \sigma_c^2, f$ across environments—the dependence of μ_e, σ_e on the environment allows for varying distributions, while the true causal process remains unchanged. Here we make a few clarifying remarks:

- We do not impose any constraints on the model parameters—in particular, we do not assume a prior. Observe that μ_c, σ_c^2 are the same for all environments, hence the subscript indicates the invariant (causal) relationship. In contrast, with some abuse of notation, the environmental subscript is used to indicate both dependence on the environment and the index of the environment itself (e.g., μ_i represents the mean specific to environment i).
- While we have framed the model as y causing z_c, the causation can just as easily be viewed in the other direction. The log-odds of y are a linear function of z_c—this matches logistic regression with an invariant regression vector $\beta_c = 2\mu_c/\sigma_c^2$ and bias $\beta_0 = \log \frac{\eta}{1-\eta}$. We present the model as above to emphasize that the causal relationships between y and the z_c, z_e are a priori indistinguishable, and because we believe this direction is more intuitive.
- The assumption of a constant marginal is necessary—in logistic regression, the optimal bias term is $\log \frac{p(y=1)}{p(y=-1)}$. Thus the optimal classifier can only be invariant if this value is the same for all environments.
- Modeling class-conditional means as direct opposites is made for clarity: it matches existing models and it greatly simplifies the calculations and results. None of our proofs require this condition, and it is straightforward to extend our results to the general case of arbitrary means.

We consider the setting where we are given infinite samples from each environment; this allows us to isolate the behavior of the objectives themselves, rather than finite-sample effects. Upon observing samples from this model, our objective is thus to learn a feature embedder Φ and regression vector $\hat{\beta}$ to minimize the risk on an unseen environment e:

$$\mathcal{R}^e(\Phi, \hat{\beta}) := \mathbb{E}_{(x,y)\sim p^e}\left[\ell(\sigma(\hat{\beta}^T \Phi(x)), y)\right].$$

The function ℓ can be any loss appropriate to classification: in this work we consider the logistic and the 0-1 loss. Note, however, that we are *not* hoping to minimize this risk in expectation over the environments; this is already accomplished via ERM or distributionally robust optimization (DRO; Bagnell 2005; Ben-Tal et al. 2009). Rather, we hope to extract and regress on invariant features while ignoring environmental features, such that our classifier generalizes to all unseen environments regardless of their parameters. In other words, the focus is on minimizing risk for the worst-case environment. We refer to the classifier which will minimize worst-case risk under arbitrary distribution shift as the *invariant classifier*. To discuss this formally, we define precisely what we mean by this term.

Definition 1. Assume the existence of a true set of "invariant features", whose joint or conditional distribution with the target variable is unchanging across environments. Then, the *invariant classifier* consists of: (a) a feature embedder which recovers exactly these features, and (b) the optimal predictor on top of these features.

This definition is intentionally broad; in this paper, we consider logistic regression and thus we assume the optimal predictor $\hat{\beta}$ is a vector—note that Φ may be non-linear. Thus, the invariant classifier comprises:

$$\Phi(x) = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \circ f^{-1}(x) = [z_c], \quad \hat{\beta} = \begin{bmatrix} \beta_c \\ \beta_0 \end{bmatrix} := \begin{bmatrix} 2\mu_c/\sigma_c^2 \\ \log \frac{\eta}{1-\eta} \end{bmatrix}.$$

[1]Note the deliberate choice to have z_e depend on y. Much work on this problem models spurious features which correlate with the label but are not causal. However, the term "spurious" is often applied incongruously; historically, a spurious correlation is one which (a) appears by chance and would disappear given enough samples or (b) is due to an unobserved confounder. In recent work, the term has been co-opted to refer to any feature that correlates with the label but does not cause it. Thus there is a subtle distinction: if we allow for *the label to cause the features* (e.g. as in natural images), the resulting correlation is *not* spurious. We therefore avoid using the term "spurious" in this work, opting instead for "non-invariant" or "environmental".

This classifier is Bayes with respect to the invariant features and so it achieves the minimum possible risk without using environmental features—it is therefore also minimax optimal. Observe that *the invariant classifier is distinct from the Bayes classifier*. The Bayes classifier uses informative but non-invariant features; the invariant classifier does not.

3.1 Informal Results

With the model defined, we can informally present our results; we defer the formal statements to first give a background on the IRM objective in the next section. For the full theorem statements, see Section 5 (linear) and Section 6 (non-linear). With a slight abuse of notation, we identify a classifier by the tuple $\Phi, \hat{\beta}$ which parametrizes it.

First, we show that the usefulness of IRM exhibits a "threshold" behavior, depending on whether $E > d_e$ or $E \leq d_e$:

Theorem 3.1 (Informal, Linear)**.** *For linear f, consider solving the IRM objective to learn a linear Φ with invariant optimal coefficients $\hat{\beta}$. If $E > d_e$, then $\Phi, \hat{\beta}$ is precisely the invariant classifier; it uses only invariant features and generalizes to all environments with minimax-optimal risk. If $E \leq d_e$, then $\Phi, \hat{\beta}$ relies upon non-invariant features.*

In fact, when $E \leq d_e$ it is even possible to learn a classifier *solely* relying on environmental features that achieves lower risk on the training environments than the invariant classifier:

Theorem 3.2 (Informal, Linear)**.** *For linear f and $E \leq d_e$, under mild conditions there exists a linear classifier $\Phi, \hat{\beta}$ which uses* only environmental features*, yet achieves lower risk than the optimal invariant classifier.*

Finally, in the non-linear case, we show that IRM fails unless the training environments approximately "cover" the space of possible environments, and therefore behaves similarly to ERM:

Theorem 3.3 (Informal, Non-linear)**.** *For non-linear f, there exists a non-linear classifier $\Phi, \hat{\beta}$ which is nearly optimal under the penalized objective and furthermore is nearly identical to the invariant classifier on the training distribution. However, for any test environment with a mean sufficiently different from the training means, this classifier will be equivalent to the ERM solution on nearly all test points. For test distributions where the environmental feature correlations with the label are reversed, this classifier has almost trivial performance.*

Extensions to other objectives. Several follow-up works have suggested alternatives to IRM (see Section 4). Though these objectives perform better on various baselines, there are few formal guarantees and no results beyond the linear case. Due to their collective similarities, it is straightforward to extend every theorem in this paper to these objectives, demonstrating that they all suffer from the same shortcomings. Appendix E contains example corollaries for each of the main results presented in this work.

4 Background on Invariant Risk Minimization and its Alternatives

During training, a classifier will learn to leverage correlations between features and labels in the training data to make its predictions. If a correlation varies with the environment, it may not be present in future test distributions—worse yet, it may be *reversed*—harming the classifier's predictive ability. IRM (Arjovsky et al., 2019) is a recently proposed approach to learning environmentally invariant representations to facilitate invariant prediction.

The IRM objective. IRM posits the existence of a feature embedder Φ such that the optimal classifier on top of these features is the same for every environment. The authors argue that such a function will use only invariant features, since non-invariant features will have different conditional distributions with the label and therefore a fixed linear classifier on top of them won't be optimal in all environments. To learn this featurizer, the IRM objective is the following constrained optimization problem:

$$\min_{\Phi, \hat{\beta}} \quad \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathcal{R}^e(\Phi, \hat{\beta}) \qquad \text{s.t.} \quad \hat{\beta} \in \arg\min_{\beta} \mathcal{R}^e(\Phi, \beta) \quad \forall e \in \mathcal{E}. \tag{4}$$

This bilevel program is highly non-convex and difficult to solve. To find an approximate solution, the authors consider a Langrangian form, whereby the sub-optimality with respect to the constraint is expressed as the squared norm of the gradients of each of the inner optimization problems:

$$\min_{\Phi, \hat{\beta}} \quad \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left[\mathcal{R}^e(\Phi, \hat{\beta}) + \lambda \|\nabla_{\hat{\beta}} \mathcal{R}^e(\Phi, \hat{\beta})\|_2^2 \right]. \tag{5}$$

Assuming the inner optimization problem is convex, achieving feasibility is equivalent to the penalty term being equal to 0. Thus, Equations 4 and 5 are equivalent if we set $\lambda = \infty$.

Alternative objectives. IRM is motivated by the existence of a featurizer Φ such that $\mathbb{E}[y \mid \Phi(x)]$ is invariant. Follow-up works have proposed variations on this objective, based instead on the strictly stronger desideratum of the invariance of $p(y \mid \Phi(x))$. Krueger et al. (2020) suggest penalizing the variance of the risks, while Xie et al. (2020) give the same objective but taking the square root of the variance. Many papers have suggested similar alternatives (Jin et al., 2020; Mahajan et al., 2020; Bellot & van der Schaar, 2020). These objectives are compelling—it seems natural to expect that solving them will yield the desired invariance. Indeed, it is easy to show that the invariant classifier constitutes a stationary point of each of these objectives:

Proposition 4.1. *Suppose the observed data are generated according to Equations 1-3. Then recovering the (parametrized) invariant classifier $\Phi(x) = [z_c]$ and $\hat{\beta} = [\beta_c, \beta_0]$ is a stationary point for Equation 4.*

The stationarity of the invariant classifier for the other objectives is a trivial corollary. However, in the following sections we will demonstrate that such a result is misleading and that a more careful investigation is necessary.

5 The Difficulties of IRM in the Linear Regime

In their work proposing IRM, Arjovsky et al. (2019) present specific conditions for an upper bound on the number of training environments needed such that a feasible linear featurizer Φ will have an invariant optimal regression vector $\hat{\beta}$. Our first result is similar in spirit but presents a substantially stronger (and simplified) upper bound in the classification setting, along with a matching lower bound: we demonstrate that observing a large number of environments—linear in the number of environmental features—is *necessary* for generalization in the linear regime.

Theorem 5.1 (Linear case)**.** *Assume f is linear. Suppose we observe E training environments. Then the following hold:*

1. *Suppose $E > d_e$. Under mild non-degeneracy conditions, any linear featurizer Φ with an invariant optimal regression vector $\hat{\beta}$ uses only invariant features, and it therefore has identical risk on all possible environments.*
2. *If $E \leq d_e$ and the environmental means μ_e are linearly independent, then there exists a linear Φ with $\text{rank}(\Phi) = d_c + d_e + 1 - E$ whose output depends on the environmental features, yet the optimal classifier on top of Φ is invariant. Further, both the logistic and 0-1 risks of this Φ and its corresponding $\hat{\beta}$ are strictly lower than those of the invariant classifier.*

Similar to Arjovsky et al. (2019), the set of environments which do not satisfy Theorem 5.1 has measure zero under any absolutely continuous density over environmental parameters. Further details, and the full proof, can be found in Appendix C.1. Since the invariant classifier is Bayes with respect to the invariant features, by the data-processing inequality the only way a classifier can achieve lower risk is by relying on environmental features. Thus, Theorem 5.1 directly implies that when $E \leq d_e$, the global minimum necessarily uses these non-invariant features and therefore will not universally generalize to unseen environments. On the other hand, in the (perhaps unlikely) case that $E > d_e$, any feasible solution will generalize, and the invariant classifier has the minimum (and minimax) risk of all such classifiers:

Corollary 5.2. *For both logistic and 0-1 loss, the invariant classifier is a global minimum of the IRM objective if and only if $E > d_e$.*

Let us compare our theoretical findings to those of Arjovsky et al. (2019). Suppose the observations x lie in $\mathbb{R}^d$. Roughly, their theorem says that for a learned Φ of rank r with invariant optimal coefficient $\hat{\beta}$, if the training set contains $d - r + d/r$ "non-degenerate" environments, then $\hat{\beta}$ will be optimal for all environments. There are several important issues with this result: first, they present no result tying the rank of Φ to their actual objective; their theory thus motivates the objective, but does not provide any performance guarantees for its solution. Next, observe when x is high-dimensional (i.e. $d \gg d_e + d_c$)—in which case Φ will be low-rank (i.e. $r \leq d_e + d_c$)—their result requires $\Omega(d)$ environments, which is extreme. For example, think of images on a low-dimensional manifold embedded in very high-dimensional space. Even when $d = d_c + d_e$, the "ideal" Φ which recovers precisely z_c would have rank d_c, and therefore their condition for invariance would require $E > d_e + d_e/d_c$, a stronger requirement than ours; this inequality also seems unlikely to hold in most real-world settings. Finally, they give no lower bound on the number of required environments—prior to this work, there were no existing results for the performance of the IRM objective when their conditions are not met. We also run a simple synthetic experiment to verify our theoretical results, drawing samples according to our model and learning a classifier with the IRM objective. Details and results of this experiment can be found in Appendix C.2.

We now sketch a proof of part 2 of the theorem for when $E = d_e$; this involves an explicit construction of a Φ of rank $d_c + 1$ whose optimal $\hat{\beta}$ is invariant:

Proof Sketch. Since f has an inverse over its range, we can define Φ as a linear function directly over the latents $[z_c, z_e]$. Specifically, we define $\Phi(x) = [z_c, p^T z_e]$. Here, p is a unit-norm vector such that $\forall e \in \mathcal{E},\ p^T \mu_e = \sigma_e^2 \tilde{\mu}$, where $\tilde{\mu}$ is a fixed scalar that depends on the geometry of the μ_e, σ_e^2—such a vector exists so long as the means are linearly independent. Observe that this Φ also has the desired rank. Since this is a linear function of a multivariate Gaussian, the label-conditional distribution of each environment's non-invariant latents has a simple closed form: $p^T z_e \mid y \sim \mathcal{N}(y \cdot p^T \mu_e, \|p\|_2^2 \sigma_e^2) \stackrel{d}{=} \mathcal{N}(y \cdot \sigma_e^2 \tilde{\mu}, \sigma_e^2)$.

For separating two Gaussians, the optimal linear classifier weights are $\Sigma^{-1}(\mu_1 - \mu_0)$—here, the optimal classifier weight on $p^T z_e$ is precisely $2\tilde{\mu}$, which does not depend on the environment (and neither do the optimal coefficients for z_c). Though the distribution varies across environments, the optimal classifier is the same! Thus, Φ directly depends on the environmental features, yet the optimal regression vector $\hat{\beta}$ for each environment is constant. To see that it has lower risk than the invariant classifier, note that this classifier is Bayes with respect to its features and that the invariant classifier uses a strict subset of these features, and therefore it has less information for its predictions. □

A purely environmental classifier. The precise value of $\tilde{\mu}$ in the proof sketch above represents how strongly this non-invariant feature is correlated with the label. In theory, a classifier that achieves a lower objective value could do so by a very small margin—incorporating an arbitrarily small amount of information from a non-invariant feature would suffice. This result would be less surprising, since achieving low empirical risk might still ensure that we are "close" to the invariant classifier. Our next result shows that this is not the case: there exists a feasible solution which uses *only the environmental features* yet performs better than the invariant classifier on all environments for which $\tilde{\mu}$ is large enough.

Theorem 5.3. *Suppose we observe $E \leq d_e$ environments, such that all environmental means are linearly independent. Then there exists a feasible $\Phi, \hat{\beta}$ which uses* only environmental features *and achieves lower 0-1 risk than the invariant classifier on every environment e such that $\sigma_e \tilde{\mu} > \sigma_c^{-1} \|\mu_c\|_2$ and $2\sigma_e \tilde{\mu} \sigma_c^{-1} \|\mu_c\|_2 \geq |\beta_0|$.*

The latter of these two conditions is effectively trivial, requiring only a small separation of the means and balance in class labels. From the construction of $\tilde{\mu}$ in the proof of Lemma C.1, we can see that the former condition is more likely to be met when $E \ll d_e$ and in environments where some non-invariant features are reasonably correlated with the label—both of which can be expected to hold in the high-dimensional setting. Figure C.2 in the appendix plots the results for a few toy examples for various dimensionalities and variances to see how often this condition holds in practice. For all settings, the number of environments observed before the condition ceases to hold is quite high—on the order of $d_e - d_c$.

6 The Failure of IRM in the Non-Linear Regime

We've demonstrated that OOD generalization is quite difficult in the linear case, but it is achievable given enough training environments. Our results—and those of Arjovsky et al. (2019)—intuitively proceed by observing that each environment reduces a "degree of freedom" of the invariant solution, such that only the invariant features remain feasible if enough environments are seen. In contrast, in the non-linear case, it's not clear how to capture this idea of restricting the "degrees of freedom"—and in fact our results imply that this intuition is simply wrong. Instead, we show that the solution generalizes only to test environments that are sufficiently similar to the training environments. Thus, these objectives present no real improvement over ERM or DRO.

Non-linear transformations of the latent variables z_c, z_e make it hard to directly characterize the optimal linear classifier $\hat{\beta}$, which makes reasoning about the constrained solution to Equation 4 difficult. Instead, we turn our attention to Equation 5, the penalized IRM objective used in practice. By analyzing this objective we can reason about solutions that are sufficiently close to optimal, which allows for concrete results on the resulting classifier. In this section we demonstrate a foundational flaw of IRM and related objectives in the non-linear regime: unless we observe enough environments to "cover" the space of non-invariant features, a solution that appears to be invariant can still wildly underperform on a new test distribution. We begin with a definition about the optimality of a coefficient vector $\hat{\beta}$:

Definition 2. For $0 < \gamma < 1$, a coefficient vector $\hat{\beta}$ is γ-*close to optimal* for a label-conditional feature distribution $z \sim \mathcal{N}(y \cdot \mu, \Sigma)$ if

$$\hat{\beta}^T \mu \geq (1 - \gamma) 2\mu^T \Sigma^{-1} \mu.$$

Since the optimal coefficient vector is precisely $2\Sigma^{-1}\mu$, being γ-close implies that $\hat{\beta}$ is reasonably aligned with that optimum. Observe that the definition does not account for magnitude—the set of vectors which is γ-close to optimal is therefore a halfspace which is normal to the optimal vector. One of our results in the non-linear case uses the following assumption, which says that the observed environmental means are sufficiently similar to one another.

Assumption 1. There exists a $0 \leq \gamma < 1$ such that the ERM-optimal coefficients for the non-invariant features,

$$\beta_{e;\text{ERM}} := \arg\min_{\hat{\beta}_e} \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathbb{E}_{z_c, z_e, y \sim p^e} \left[\ell(\sigma(\beta_c^T z_c + \hat{\beta}_e^T z_e + \beta_0), y) \right], \tag{6}$$

is γ-close to optimal for every environmental feature distribution in $\mathcal{E}$.

This assumption says the environmental distributions are similar enough such that the optimal "average classifier" is reasonably predictive for each environment individually. This is a natural expectation: we are employing IRM precisely *because* we expect the ERM classifier to do well on the training set but fail to generalize. If the environmental parameters are sufficiently orthogonal, we might instead expect ERM to ignore the features which are not at least moderately predictive across all environments. Finally, we note that if this assumption only holds for a subset of features, our result still applies by marginalizing out the dimensions for which it does not hold.

We are now ready to give our main result in the non-linear regime. We present a simplified version, assuming that that $\sigma_e^2 = 1 \; \forall e$. This is purely for clarity of presentation; the full theorem, for which the result still holds, is presented in Appendix D. We make use of two constants in the following proof—the average squared norm of the environmental means, $\overline{\|\mu\|_2^2} := \frac{1}{E} \sum_{e \in \mathcal{E}} \|\mu_e\|_2^2$; and the standard deviation of the response variable of the ERM-optimal classifier, $\sigma_{\text{ERM}} := \sqrt{\|\beta_c\|_2^2 \sigma_c^2 + \|\beta_{e;\text{ERM}}\|_2^2 \sigma_e^2}$.

Theorem 6.1 (Non-linear case, simplified)**.** *Suppose we observe E environments $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$, where $\sigma_e^2 = 1 \; \forall e$. Then, for any $\epsilon > 1$, there exists a featurizer Φ_ϵ which, combined with the ERM-optimal classifier $\hat{\beta} = [\beta_c, \beta_{e;ERM}, \beta_0]^T$, satisfies the following properties, where we define $p_{\epsilon,d_e} := \exp\{-d_e \min((\epsilon - 1), (\epsilon - 1)^2)/8\}$:*

1. *The regularization term of $\Phi_\epsilon, \hat{\beta}$ as in Equation 5 is bounded as*

$$\frac{1}{E} \sum_{e \in \mathcal{E}} \|\nabla_{\hat{\beta}} \mathcal{R}^e(\Phi_\epsilon, \hat{\beta})\|_2^2 \in \mathcal{O}\left(p_{\epsilon,d_e}^2 \left(c_\epsilon d_e + \overline{\|\mu\|_2^2} \right) \right),$$

   for some constant c_ϵ that depends only on ϵ.

2. *$\Phi_\epsilon, \hat{\beta}$ exactly matches the invariant classifier on at least a $1 - p_{\epsilon,d_e}$ fraction of the training set. On the remaining inputs, it matches the ERM-optimal solution.*

Further, for any test distribution, suppose its environmental mean μ_{E+1} is sufficiently far from the training means:

$$\forall e \in \mathcal{E}, \min_{y \in \{\pm 1\}} \|\mu_{E+1} - y \cdot \mu_e\|_2 \geq (\sqrt{\epsilon} + \delta)\sqrt{d_e} \tag{7}$$

for some $\delta > 0$, and define $q := \frac{2E}{\sqrt{\pi}\delta} \exp\{-\delta^2\}$. Then the following holds:

3. *$\Phi_\epsilon, \hat{\beta}$ is equivalent to the ERM-optimal classifier on at least a $1 - q$ fraction of the test distribution.*

4. *Under Assumption 1, suppose it holds that $\mu_{E+1} = -\sum_{e \in \mathcal{E}} \alpha_e \mu_e$ for some set of coefficients $\{\alpha_e\}_{e \in \mathcal{E}}$. Then so long as*

$$\sum_{e \in \mathcal{E}} \alpha_e \|\mu_e\|_2^2 \geq \frac{\|\mu_c\|_2^2 / \sigma_c^2 + |\beta_0|/2 + \sigma_{ERM}}{1 - \gamma}, \tag{8}$$

   the 0-1 risk of $\Phi_\epsilon, \hat{\beta}$ on the new environment is greater than $.975 - q$.

Before giving a proof sketch, we give a brief intuition for each of the claims made in this theorem:

1. The first claim says that the classifier we construct will have a gradient squared norm scaling as p_{ϵ,d_e}^2 which is exponentially small in d_e. Thus, in high dimensions, the penalty term for our classifier will shrink rapidly towards 0, meaning it will appear as a perfectly reasonable solution to the objective (5).
2. The second claim says that this "fake" invariant classifier is identical to the true one on all but an exponentially small fraction of the training data; on the remaining fraction, it matches the ERM-optimal solution, which has lower risk. The correspondence between constrained and penalized optimization implies that for large enough d_e, the "fake" classifier will often be a preferred solution. In the finite-sample setting, we would need exponentially many samples to even distinguish between the two!

3. The third claim is the crux of the theorem; it says that this classifier we've constructed will completely fail to use invariant prediction on most environments. Recall, the intent of IRM is to perform well precisely when ERM breaks down: when the test distribution differs greatly from the training distribution. If we assume a Gaussian prior on the training environment means, they will have an ℓ_2-norm separation in $\mathcal{O}(\sqrt{d_e})$ with high probability. Observe that q will be vanishingly small so long as $\delta \geq \text{polylog}(E)$. Thus, part 3 says that IRM fails to use invariant prediction on any environment that is even slightly outside the high probability region of the training prior; even a separation of $\Omega(\sqrt{d_e \log E})$ suffices. If we instead expect the new environments to be similar, ERM already guarantees reasonable average performance at test-time; thus, IRM fundamentally *does not improve* over ERM in the non-linear regime.
4. The final statement demonstrates a particularly egregious failure case of this classifier: just like ERM, if the correlation between the non-invariant features and the label reverses at test-time, our classifier will have significantly worse than chance performance.

A back-of-the-envelope calculation for Equations 7, 8 and Assumption 1 makes clear just how broadly we can expect them to hold; for details, see Appendix D.2. With this intuition, we now present a sketch of the proof—the full proof can be found in Appendix D.

Proof Sketch. The key is a construction of Φ which is almost identical to the invariant classifier on the observed environments, yet behaves like the ERM solution at test time. We partition the environmental feature space into two sets, $\mathcal{B}, \mathcal{B}^c \subset \mathbb{R}^{d_e}$. $\mathcal{B}$ is the union of balls centered at each μ_e, each with a large enough radius that it contains the majority of samples from that environment; therefore $\mathcal{B}$ represents the vast majority of the training distribution. On this set, define $\Phi(x) = [z_c]$, so our construction is equal to the invariant classifier. Now consider $\mathcal{B}^c = \mathbb{R}^{d_e} \setminus \mathcal{B}$. We use standard concentration results to upper bound the measure of $\mathcal{B}^c$ under the training distribution by p_{ϵ,d_e}. Next, we show how choosing $\Phi(x) = f^{-1}(x) = [z_c, z_e]^T$ on this set results in the gradient squared norm sub-optimality bound, which is of order p^2_{ϵ,d_e} (part 1). It is also clear that our constructed classifier is equivalent to the ERM-optimal solution on $\mathcal{B}^c$ (part 2). Thus, our classifier will have often have lower empirical risk on $\mathcal{B}^c$, counter-weighting the regularization penalty.

The second part of the proof is to demonstrate that while $\mathcal{B}$ has large measure under the training environments, it will have very small measure (upper bounded by q) under any moderately different test environment. We can see this by considering the separation of means (Equation 7); the measure of each ball in $\mathcal{B}$ can be bounded by the measure of the halfspace containing it; if each ball is far enough away from μ_{E+1}, then the total measure of these halfspaces must be small. At test time, our classifier will therefore match the ERM solution on all but q of the observations (part 3). Finally, we lower bound the 0-1 risk of the ERM classifier under such a distribution shift by analyzing the distribution of the response variable. The proof is completed by observing that our classifier's risk can differ from this by at most q. □

Theorem 6.1 shows that it's possible for the IRM solution to perform poorly on environments which differ even moderately from the training data. We can of course guarantee generalization if the training distributions "cover" (or approximately cover) the full space of environments in order to tie down the performance on future distributions. But in such a scenario, there would no longer be a need for ICP; we could expect ERM or DRO to perform just as well. Once more, we find that our result trivially extends to the alternative objectives; we again refer to Appendix E.

7 Conclusion

Out-of-distribution generalization is an important direction for future research, and Invariant Causal Prediction remains a promising approach. However, formal results for latent variable models are lacking, particularly in the non-linear setting. This paper demonstrates that Invariant Risk Minimization and subsequent works have significant under-explored risks and issues with their formulation. In the linear regime, a large number of environments is *necessary* to ensure recovery of the invariant classifier; the number of training environments needed to ensure good generalization under arbitrary distribution shift is linear in the number of non-invariant features, which is an unrealistic expectation for most real-world distributions. In the non-linear setting, these objectives break down even further: this work presents the first such result, formally proving that the intuition behind linear risk extrapolation *does not extend to the non-linear setting*. We find that IRM suffers from precisely the same weakness as ERM, failing to generalize unless the training environments sufficiently "cover" the space of non-invariant features. This raises the question: what is the correct formulation for invariant prediction when the observations are complex, non-linear functions of unobserved latent factors? For future work, it would be helpful to investigate ours or similar models further; some interesting directions include (a) characterizing in which settings specifically an "invariance-like" constraint can lead to improved performance over ERM and (b) formulating an objective that encourages invariance in the non-linear setting in a way that can be formally demonstrated and quantified.

REFERENCES

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

J Andrew Bagnell. Robust supervised learning. In *Proceedings of the 20th national conference on Artificial intelligence-Volume 2*, pp. 714–719, 2005.

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

Alexis Bellot and Mihaela van der Schaar. Generalization and invariances in the presence of unobserved confounding. *arXiv preprint arXiv:2007.10653*, 2020.

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.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

Kenneth A Bollen. Structural equation models. *Encyclopedia of biostatistics*, 7, 2005.

Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *arXiv preprint arXiv:2006.07433*, 2020.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. In *Advances in Neural Information Processing Systems*, pp. 3011–3021, 2017.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pp. 2839–2848, 2016.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 125–136. Curran Associates, Inc., 2019.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Domain extrapolation via regret minimization. *arXiv preprint arXiv:2006.03908*, 2020.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. *arXiv preprint arXiv:1903.03448*, 2019.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 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.

Frank R Kschischang. The complementary error function. 2017. URL https://www.comm.utoronto.ca/frank/notes/erfc.pdf.

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

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018.

Marloes H Maathuis, Markus Kalisch, Peter Bühlmann, et al. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

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

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. volume 28 of *Proceedings of Machine Learning Research*, pp. 10–18, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/muandet13.html.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

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

J Peters, D Janzing, and B Schölkopf. Elements of causal inference-foundations and learning algorithms. 2017.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society*, 2016.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 5014–5026. Curran Associates, Inc., 2018.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3118–3127. PMLR, 2019.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in neural information processing systems*, pp. 5334–5344, 2018.

Chuanlong Xie, Fei Chen, Yue Liu, and Zhenguo Li. Risk variance penalization: From distributional robustness to causality. *arXiv preprint arXiv:2006.07544*, 2020.

Kun Zhang, Mingming Gong, and Bernhard Scholkopf. Multi-source domain adaptation: a causal view. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 3150–3157, 2015.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532, 2019.