MALIGN OVERFITTING: INTERPOLATION AND INVARIANCE ARE FUNDAMENTALLY AT ODDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned classifiers should often possess certain invariance properties meant to encourage fairness, robustness, or out-of-distribution generalization. However, multiple recent works empirically demonstrate that common invariance-inducing regularizers are ineffective in the over-parameterized regime, in which classifiers perfectly fit (i.e. interpolate) the training data. This suggests that the phenomenon of "benign overfitting", in which models generalize well despite interpolating, might not favorably extend to settings in which robustness or fairness are desirable.

In this work we provide a theoretical justification for these observations. We prove that - even in the simplest of settings - any interpolating classifier (with nonzero margin) will not satisfy these invariance properties. We then propose and analyze an algorithm that - in the same setting - successfully learns a non-interpolating classifier that is provably invariant. We validate our theoretical observations regarding the conflict between interpolation and invariance on simulated data and the Waterbirds dataset.

1 INTRODUCTION

Modern machine learning applications often call for models which are not only accurate, but which are also robust to distribution shifts and satisfy fairness constraints. For example, we might wish to avoid using hospital specific traces in X-ray images (DeGrave et al., 2021; Zech et al., 2018), as they rely on spurious correlations that will fail when deployed in a new hospital, or we might seek models with similar error rates across protected demographic groups in the context of loan applications (Byanjankar et al., 2015). A developing paradigm for fulfilling such requirements is learning models that satisfy some notion of *invariance* (Peters et al., 2016; 2017) across environments or sub-populations. For example, in the X-ray case, spurious correlations can be formalized as relationships between a feature and a label which are non-invariant across hospitals (Zech et al., 2018). Equal Opportunity (Hardt et al., 2016) can be expressed as a statistical constraint on the outputs of the model, where the false negative rate is invariant to membership in a protected group. Many techniques for learning invariant models have been proposed including penalties that encourage notions of invariance (Arjovsky et al., 2019; Krueger et al., 2020; Veitch et al., 2021; Wald et al., 2021; Puli et al., 2021; Makar et al., 2021; Rame et al., 2022; Kaur et al., 2022), data re-weighting (Sagawa et al., 2019; Wang et al., 2021; Idrissi et al., 2022), causal graph analysis (Subbaswamy et al., 2019; 2022), and more (Ahuja et al., 2020).

While the invariance paradigm holds promise for delivering robust or fair models, many current invariance-inducing methods often fail to improve over naive approaches. This is especially noticeable when these methods are used with overparameterized deep models capable of *interpolating*, i.e., capable of perfectly fitting the training data (Gulrajani & Lopez-Paz, 2020; Dranker et al., 2021; Guo et al., 2022; Zhou et al., 2022; Menon et al., 2021; Veldanda et al., 2022; Cherepanova et al., 2021). Existing theory explains why overparameterization hurts invariance for standard interpolating learning rules, such as empirical risk minimization and max-margin classification Sagawa et al. (2020); Nagarajan et al. (2020); D'Amour et al. (2020), and also why reweighting and some types of distributionally robust optimization face challenges when used with overparameterized models Byrd & Lipton (2019); Sagawa et al. (2019).

Researchers have responded to this situation with two parallel lines of research. The first line is attempts to come up with alternative learning rules that are capable of interpolating while still endowing meaningful invariance properties to the solutions they find (Kini et al., 2021; Wang et al., 2021). These works are motivated in part by the phenomenon of "benign overfitting" (Belkin, 2021; Bartlett et al., 2020), whereby interpolating overparameterized models achieve excellent generalization performance on an identically-distributed test set (Cao et al., 2021; Wang & Thrampoulidis, 2021; Shamir, 2022); by looking for interpolating and invariant learning rules, researchers hope to extend benign overfitting to out-of-distribution test sets. The second line of research forgoes interpolation, and instead applies inducing techniques with small models on top of representations learned by some other means (Rosenfeld et al., 2022; Veldanda et al., 2022; Kirichenko et al., 2022; Menon et al., 2021; Kumar et al., 2022; Zhang et al., 2022), as well as by subsampling techniques (Idrissi et al., 2022; Chatterji et al., 2022). As both lines of research report encouraging empirical results, it is not clear which one is the preferred way forward.

In this work we theoretically address this question, showing that interpolating models are fundamentally less invariant than non-interpolating ones. In other words, beyond identically-distributed test sets, overfitting is no longer benign. More concretely, we consider linear classification in a basic overparameterized Gaussian mixture model with invariant "core" features as well as environment-dependent "spurious" features, similar to models used in previous work to gain insight into robustness and invariance (Schmidt et al., 2018; Rosenfeld et al., 2020; Sagawa et al., 2020). We show that any learning rule producing a classifier that separates the data with non-zero margin must necessarily rely on the spurious features in the data, and therefore cannot be invariant. Moreover, in the same setting we analyze a simple two-stage algorithm that can find nearly invariant linear classifiers, i.e., with almost no dependence on the spurious. Thus, we establish a separation between the invariance attained by interpolating and non-interpolating learning rules. We believe that learning rules which fail in the simple over-parameterized linear classification setting we consider are not likely to succeed in more complicated, real-world settings. Therefore, our analysis provides useful guidance for future research into robust and fair machine learning models.

We formally state our full result in Theorem 1. In Section 3 we outline the arguments leading to the negative part of Theorem 1, i.e., the failure of interpolating classifiers to be invariant in our model. In Section 4 we establish the positive part Theorem 1, by providing and analyzing an non-interpolating algorithm that, in our model, achieves low robust error. We validate our theoretical findings with simulations and experiments on the Waterbirds dataset in Section 5, and conclude with a discussion of additional related results and directions for future research in Section 6.

2 Overview of Setting and Results

Our analysis focuses on learning linear models over data collected from a mixture of two Gaussian distributions as follows.

Definition 1. *An* environment *is a distribution parameterized by* $(\boldsymbol{\mu}_c, \boldsymbol{\mu}_s, d, \sigma, \theta)$ *where* $\theta \in [-1, 1]$ *and* $\boldsymbol{\mu}_c, \boldsymbol{\mu}_s \in \mathbb{R}^d$ *satisfy* $\boldsymbol{\mu}_c \perp \boldsymbol{\mu}_s$ *and with samples generated according to:*

$$\mathbb{P}_\theta(y) = \text{Unif}\{-1, 1\}, \quad \mathbb{P}_\theta(\mathbf{x}|y) = \mathcal{N}(y\boldsymbol{\mu}_c + y\theta\boldsymbol{\mu}_s, \sigma^2 I). \tag{1}$$

We focus on problems with two environments that share all their parameters other than θ, hence we will have $(\boldsymbol{\mu}_c, \boldsymbol{\mu}_s, d, \sigma, \theta_e)$ for $e \in \{1, 2\}$, as we define next:

Definition 2 (Linear Two Environment Problem)**.** *In a Linear Two Environment Problem we have datasets* S_1, S_2 *of sizes* N_1, N_2 *drawn from* $\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}$ *respectively, where* $\boldsymbol{\mu}_c$ *and* $\boldsymbol{\mu}_s$ *satisfy* $\|\boldsymbol{\mu}_c\| = r_c$ *and* $\|\boldsymbol{\mu}_s\| = r_s$*. A learning algorithm is a (possibly randomized) mapping from the tuple* (S_1, S_2) *to a linear classifier* $\mathbf{w} \in \mathbb{R}^d$ *and we refer to* $S_1 \cup S_2$ *as the pooled dataset* $S = \{\mathbf{x}_i, y_i\}_{i=1}^N$*, where* $N = N_1 + N_2$.

We study settings where θ_1, θ_2 are fixed and d is large compared to N, i.e. the overparameterized regime. We refer to the two distributions $\mathbb{P}_{\theta_e}$ for $e \in \{1, 2\}$ as "training environments", following Peters et al. (2016); Arjovsky et al. (2019). In the context of Out-of-Distribution (OOD) generalization, environments correspond to different experimental conditions, e.g. collection of medical data in two hospitals. In a fairness context, we may think of these distributions as subpopulations (e.g.

demographic groups)[1]. While these are different applications that require specialized methods, the underlying formalism of solutions is often similar (see, e.g., Creager et al., 2021, Table 1), where we wish to learn a classifier that in one way or another is invariant to the environment variable.

The power of the simple model defined in Definition 2 is that many of the common invariance criteria all boil down to the same mathematical constraint: learning a classifier that is orthogonal to $\boldsymbol{\mu}_s$, which induces a spurious correlation between the environment and the label. These include Equalized Odds (Hardt et al., 2016), conditional distribution matching Li et al. (2018), calibration on multiple subsets of the data (Hébert-Johnson et al., 2018; Wald et al., 2021), Risk Extrapolation (Krueger et al., 2020) and CVaR fairness (Williamson & Menon, 2019).

In terms of predictive accuracy, the goal of learning a linear model that aligns with $\boldsymbol{\mu}_c$ (the invariant part of the data generating process for the label) and is orthogonal to $\boldsymbol{\mu}_s$ coincides with providing guarantees on the robust error, i.e. the error when data is generated with values of θ that are different from the θ_1, θ_2 used to generate training data [2].

Definition 3 (Robust error)**.** *The robust error of a linear classifier with weights* $\mathbf{w} \in \mathbb{R}^d$ *is:*

$$\max_{\theta \in [-1,1]} \epsilon_\theta(\mathbf{w}), \text{ where } \epsilon_\theta(\mathbf{w}) := \mathbb{E}_{\mathbf{x}, y \sim \mathbb{P}_\theta} \left[\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \neq y\right]. \tag{2}$$

Having fully defined the setting, we may now turn to overview our main result.

2.1 Statement of Main Result

The question we study is whether algorithms that perfectly fit, i.e. interpolate, their training data can learn models with low robust error. To give a meaningful answer, we use the notion of normalized margin. Ideally we would like to give a result on all classifiers that attain training error zero in terms of the 0-1 loss. However, the inherent discontinuity of this loss would make any such statement sensitive to instabilities and pathologies. For instance, if we do not limit the capacity of our models, we can turn any classifier into an interpolating one by adding "special cases" for the training points, yet intuitively this is not the type of interpolation that we would like to study. Hence the margin serves as a continuous and smooth surrogate for this notion.

Definition 4 (Normalized margin)**.** *Let* $\gamma > 0$*, we say a classifier* $\mathbf{w} \in \mathbb{R}^d$ *separates the set* $S = \{\mathbf{x}_i, y_i\}_{i=1}^N$ *with* normalized margin γ *if it satisfies for each point in* S*:*

$$\frac{y_i \langle \mathbf{w}, \mathbf{x}_i \rangle}{\|\mathbf{w}\|} > \gamma \sqrt{\sigma^2 d}.$$

The $\sqrt{\sigma^2 d}$ scaling of γ is roughly proportional to $\|\mathbf{x}\|$ under our data model in Equation (1), and keeps the value of γ comparable across growing values of d. Equipped with this definition we now state our main result, and in the subsequent sections we will interpret and discuss its details.

Theorem 1. *For any sample sizes* $N_1, N_2 > 65$*, margin lower bound* $\gamma < \frac{1}{4\sqrt{N_1+N_2}}$ *and target robust error* $\epsilon > 0$*, there exist parameters* $r_c, r_s > 0, d > N_1 + N_2, \sigma, \theta_1, \theta_2$ *such that the following holds for the Linear Two Environment Problem (Definition 2) with these parameters.*

1. ***Invariance is attainable****. Algorithm 1 maps* (S_1, S_2) *to a linear classifier* $\mathbf{w}$ *such that with probability at least* $99/100$ *(over the draw* S*), the* robust error *of* $\mathbf{w}$ *is less than* ϵ*.*

2. ***Interpolation is attainable****. With probability at least* $99/100$*, the signed-sample-mean estimator* $\mathbf{w}_{\text{mean}} = N^{-1} \sum_{i \in [N]} y_i \mathbf{x}_i$ *separates* S *with normalized margin (Definition 4) greater than* $\frac{1}{4}(N_1 + N_2)^{-1/2}$*.*

[1] We note that in some settings, more commonly in the fairness literature, e is treated as a feature given to the classifier as input. Our focus is on cases where this is either impossible or undesired. For instance, because at test time e is unobserved or ill-defined (e.g. we obtain data from a new hospital). However, we emphasize that the *leaning rules* we consider have full knowledge of which environment produced each training example

[2] In fact, as we show in Equation (5) in Section 3, learning a model orthogonal to $\boldsymbol{\mu}_s$ is also a necessary condition to minimize the robust error. Attaining guarantees on the robust error will thus also have consequences on invariance of the model, as defined by these criteria. We discuss this further in section F of the appendix.

3. ***Interpolation is at odds with invariance.*** *Given $\boldsymbol{\mu}_c$ uniformly distributed on the sphere of radius r_c and $\boldsymbol{\mu}_s$ uniformly distributed on a sphere of radius r_s in the subspace orthogonal to $\boldsymbol{\mu}_c$, let* $\mathbf{w}$ *be any classifier learned from (S_1, S_2) as per Definition 2. If* $\mathbf{w}$ *separates S with normalized margin γ, then with probability at least $99/100$ (over the draw of $\boldsymbol{\mu}_c, \boldsymbol{\mu}_s$, and the sample), the* robust error *of* $\mathbf{w}$ *is at least $1/2$.*

Essentially, Theorem 1 shows that if a learning algorithm for overparameterized linear classifiers always separates its training data, then there exist natural settings for which the algorithm completely fails to learn a robust classifier, and will therefore fail on multiple other invariance and fairness objectives. It also shows that failure can be avoided in these settings, as there exists an algorithm (that *necessarily* does not always separate its training data) which successfully learns an invariant classifier. Since the learning problem we consider is simple, we believe that any reasonable method for learning invariant classifiers in the overparameterized regime should be expected to succeed here; the failure to do so highlights a fundamental limitation of all inpterpolating learning rules.

We establish Theorem 1 with three propositions, each corresponding to an enumerated claim in the theorem: (1) Proposition 2 (Section 4) establishes that invariance is attainable, (2) Proposition 3 (Appendix C) establishes that interpolation is attainable, and (3) Proposition 1 (in Section 3) establishes that interpolation is at odds with invariance. We choose to begin with the latter proposition since it is the main conceptual and technical contribution of our paper. Conversely, Proposition 3 is an easy byproduct of the developments leading up to Proposition 1, and we therefore defer it to the supplementary material.

With Propositions 1, 2 and 3 in hand, the proof of Theorem 1 simply consists of choosing the free parameters in Theorem 1 $r_c, r_s, d, \sigma, \theta_1$ and θ_2 based on these propositions such that all the claims in the theorem hold simultaneously. For convenience we take $\theta_1 = 1$, $\theta_2 = 0$, $\sigma^2 = 1/d$. Then (ignoring constant factors) we pick $r_s^2 \propto \frac{1}{N}$ and $r_c^2 \propto r_s^2/(1 + \frac{\sqrt{N_2}}{N_1\gamma})$ in order to satisfy requirements in Propositions 1 and 3. Finally, we take d to be sufficiently large so as to satisfy the remaining requirements; the resulting setting scales as $d \propto \max\left\{N^2, \frac{N}{\gamma^2 N_1^2 r_c^2}, \frac{(Q^{-1}(\epsilon))^2}{N_{\min} r_c^4}, \frac{1}{N_{\min}^2 r_c^4}\right\}$, where $N_{\min} = \min\{N_1, N_2\}$ and Q is the Gaussian tail function (see Appendix E).

A notable aspect of our result is that it holds for essentially all values of N_2 and N_1. This stands in contrast to prior work on barriers to robustness (e.g., Sagawa et al., 2020), which typically relies on one of the environment being under-represented, i.e., $N_2 \ll N_1$. We are able to sidestep such requirements by making the invariant signal component (r_c) much weaker than the spurious component (r_s), while still allowing for low test error by taking the problem dimension to be sufficiently high. However, when one environment is sufficiently rare (namely $N_2 \leq N_1^2\gamma^2$), we can show that interpolation precludes invariance even when r_s and r_c are of the same order.

Finally, we note that our results hold for classifiers with *arbitrarily small* margin γ, for settings where the maximum achievable margin is always at least of the order of $1/\sqrt{N_1 + N_2}$. Therefore, we believe that Theorem 1 essentially precludes any learning that always fits the data from being consistently invariant.

3 Interpolating Models Cannot Be Invariant

In this section we prove the third claim in Theorem 1. We set the parameters $\sigma^2 d = 1$ and $\theta_1 = 1, \theta_2 = 0$, meaning the spurious correlation is prevalent in the first environment and absent from the second. As stated in Theorem 1, our claim is that, for essentially any nonzero value of the normalized margin γ, there are instances of the Linear Two Environment Problem (Definition 2) where with high probability, learning algorithms that return linear classifiers attaining normalized margin at least γ must incur a large robust error. The following proposition lays out the claim from which we derive our statement in Theorem 1, its proof can be found in Appendix B.3.

Proposition 1. *There are universal constants $c_n \in (0, 1)$ and $C_d, C_r \in (1, \infty)$, such that, for any target normalized γ and failure probability $\delta \in (0, 1)$, if*

$$\max\{r_s^2, r_c^2\} \leq \frac{c_n}{N} \ , \ \frac{r_s^2}{r_c^2} \geq C_r \left(1 + \frac{\sqrt{N_2}}{N_1\gamma}\right) \quad \textit{and} \tag{3}$$

$$d \geq C_d \frac{N}{\gamma^2 N_1^2 r_c^2} \log \frac{1}{\delta}, \tag{4}$$

then with probability at least $1-\delta$ *over the drawing of* $\boldsymbol{\mu}_c, \boldsymbol{\mu}_s$ *and* (S_1, S_2) *as described in Theorem 1, any* $\hat{\mathbf{w}} \in \mathbb{R}^d$ *that is a measurable function of* (S_1, S_2) *and separates the data with normalized margin larger than* γ *has robust error at least* 0.5.

Proof sketch. The main part of the proof draws a lower bound on the ratio $\langle \mathbf{w}, \boldsymbol{\mu}_s \rangle / \langle \mathbf{w}, \boldsymbol{\mu}_c \rangle$ (with high probability) that is approximately $\left(\|\boldsymbol{\mu}_s\|^2 N_1 \gamma\right) / \left(\|\boldsymbol{\mu}_c\|^2 \sqrt{N_2}\right)$. Therefore, for a classifier that attains margin γ satisfying Equation (3), this ratio is likely to be larger than 1. The ratio directly relates to the robust error: for linear classifiers and Gaussian data, the error $\epsilon_\theta(\mathbf{w})$ is

$$\epsilon_\theta(\mathbf{w}) = Q\left(\frac{\langle \mathbf{w}, \boldsymbol{\mu}_c \rangle + \theta \langle \mathbf{w}, \boldsymbol{\mu}_s \rangle}{\sigma \|\mathbf{w}\|}\right) = Q\left(\frac{\langle \mathbf{w}, \boldsymbol{\mu}_c \rangle}{\sigma \|\mathbf{w}\|}\left(1 + \theta \frac{\langle \mathbf{w}, \boldsymbol{\mu}_s \rangle}{\langle \mathbf{w}, \boldsymbol{\mu}_c \rangle}\right)\right), \tag{5}$$

where $Q(t) \coloneqq \mathbb{P}(\mathcal{N}(0;1) > t)$ is the Gaussian tail function. Whenever $\langle \mathbf{w}, \boldsymbol{\mu}_s \rangle / \langle \mathbf{w}, \boldsymbol{\mu}_c \rangle > 1$, it is easy to see that $\epsilon_\theta(\mathbf{w}) = 1/2$ for some $\theta \in [-1, 1]$ and therefore the robust error is at least $\frac{1}{2}$.

To obtain the aforementioned lower bound, we first claim that if we fix a training set $\{\mathbf{x}_i, y_i\}_{i=1}^N$, then the component of $\mathbf{w}$ that is orthogonal to the training set has a negligible contribution to the performance of the classifier (see Corollary 1 in the appendix). This is due to the random generation of $\boldsymbol{\mu}_c, \boldsymbol{\mu}_s$ in our data generating process. Consequently we may write $\mathbf{w} \approx \sum_i \mathbf{x}_i \beta_i$ for some vector $\beta \in \mathbb{R}^N$, and inner products with $\mathbf{w}$ (e.g. $\langle \mathbf{w}, \boldsymbol{\mu}_s \rangle, \langle \mathbf{w}, \mathbf{x}_i \rangle$) can be expressed as linear functions of β. This lets us draw bounds on $\langle \mathbf{w}, \boldsymbol{\mu}_s \rangle$ and $\langle \mathbf{w}, \boldsymbol{\mu}_c \rangle$ under margin constraints via convex duals of the suitable constrained quadratic programs (see Lemma 4 in appendix). These components are put together in Appendix B.3 of the appendix to obtain the bound of interest. □

Implication for invariance-inducing algorithms. Our proof implies that any interpolating algorithm should fail at learning invariant classifiers. However, it does not discuss whether algorithms from the literature that have been proposed for this problem indeed interpolate their training data. In principle, it could be that the conditions imposed by all these algorithms avoid interpolation. Our simulations in Section 5 will show that several popular algorithms proposed for this problem indeed interpolate their data in the overparameterized regime. Our results may seem discouraging for the development of invariance inducing techniques using overparameterized models. It is natural to ask what type of methods *can* provably learn such models, which leads to the topic of the next section and the first part of Theorem 1.

4 A Provably Invariant Overparameterized Estimator

We now turn to propose and analyze an algorithm (Algorithm 1) that provably learns an overparametrized linear model with good robust accuracy in our setup. Our approach is a two-staged learning procedure that is conceptually similar to some recently proposed methods (Rosenfeld et al., 2022; Veldanda et al., 2022; Kirichenko et al., 2022; Menon et al., 2021; Kumar et al., 2022; Zhang et al., 2022). In Section 5 we validate our algorithm on simulations and on the Waterbirds dataset Sagawa et al. (2019), but we leave a thorough empirical evaluation of the techniques described here to future work.

Let us describe the operation of Algorithm 1. First, we evenly[3] split the data from each environment into the sets $S_e^{\text{trn}}, S_e^{\text{fine}}$, for $e \in \{1, 2\}$. The two stages of the algorithm operates on different splits of the data as follows.

1. **"Training" stage**: We use S_e^{trn} to fit an overparameterized, interpolating classifier $\mathbf{w}_e$ *separately* for each environment $e \in \{1, 2\}$.
2. **"Fine-tuning" (or post-processing) stage**: We use the (pooled) second portion of the data $S^{\text{fine}} = \{S_1^{\text{fine}}, S_2^{\text{fine}}\}$ to learn an invariant linear classifier over a new representation, which concatenates the outputs of the classifiers in the first stage. In particular, we learn this classifier by maximizing a score (i.e., minimizing an empirical loss), subject to an empirical version of an invariance constraint. For generality we denote this constraint by membership in some set of functions $\mathcal{F}(S^{\text{fine}})$.

[3]The even split is used here for simplicity of exposition, and our full proof does not assume it. In practice, allocating more data to the first-stage split would likely perform better.

Algorithm 1 Two Phase Learning of Overparameterized Invariant Classifiers

Input: Dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ and a partition S_1, S_2 into environments. Invariance constraint function family $\mathcal{F}(\cdot)$

Output: A classifier $f_{\mathbf{v}}(\mathbf{x})$

Draw subsets of data $S_{\text{trn}} = \cup_{e\in\{1,2\}} S_e^{\text{trn}}$, where $S_e^{\text{trn}} \subset S_e$ for $e \in \{1,2\}$ and $|S_e^{\text{trn}}| = N_e/2$

Stage 1: Calculate $\mathbf{w}_e = N_e^{-1} \sum_{i\in S_e^{\text{trn}}} \mathbf{x}_i y_i$ for each $e \in \{1,2\}$

Define $S^{\text{fine}} = S \setminus S^{\text{trn}}$

Stage 2: Return the solution $f_{\mathbf{v}}(\mathbf{x}; S_{\text{trn}}) = \langle v_1 \cdot \mathbf{w}_1 + v_2 \cdot \mathbf{w}_2, \mathbf{x}\rangle$ that solves

$$\text{maximize} \sum_{i\in S^{\text{fine}}} f_{\mathbf{v}}(\mathbf{x}_i) y_i \quad \text{subject to} \quad \|\mathbf{v}\|_\infty = 1 \quad \text{and} \quad f_{\mathbf{v}} \in \mathcal{F}(S^{\text{fine}}) \tag{7}$$

Crucially, the invariance penalty is only used in the second stage, in which we are no longer in the overparamterized regime since we are only fitting a two-dimensional classifier. In this way, we overcome the negative result from Section 3.

While our approach is general and can handle a variety of invariance notions (we discuss some of them in Appendix F), we will analyze the algorithm under the Equal Opportunity (EOpp) criterion (Hardt et al., 2016). Namely, for a model $f : \mathbb{R}^d \to \mathbb{R}$ we write:

$$\mathcal{F}(S^{\text{fine}}) = \left\{ f : \hat{T}_1(f; S^{\text{fine}}) = \hat{T}_2(f; S^{\text{fine}}) \right\}, \tag{6}$$
$$\text{where } \hat{T}_e(f; S^{\text{fine}}) := \frac{4}{N_e} \sum_{(\mathbf{x}_i, y_i)\in S_e^{\text{fine}}: y_i = 1} f(\mathbf{x}_i).$$

Note that this is the empirical version of the constraint $\mathbb{E}_{\mathbb{P}_{\theta_1}}[f(x)|y=1] = \mathbb{E}_{\mathbb{P}_{\theta_2}}[f(x)|y=1]$. From a fairness perspective (e.g. thinking of a loan application), this constraint ensures that the "qualified" members (i.e. those with $y = 1$) of each group receive similar predictions, on average over the entire group.

We now turn to providing conditions under which Algorithm 1 successfully learns an invariant predictor, followed by an empirical validation of our main results. The full proof for the following proposition can be found in section D.1 of the appendix.

Proposition 2. *Consider the Linear Two Environment Problem (Definition 2), and further suppose that $|\theta_1 - \theta_2| > 0.1$.[4] Let $\epsilon > 0, \delta \in (0,1)$ denote the target robust error of the model and failure probability of the algorithm, respectively. Let $N_{\min} = \min\{N_1, N_2\} \geq C_{\text{opp}} \log(1/\delta)$ for some $C_{\text{opp}} \in (1,\infty)$[5] and assume that for some constants $C_c, C_s \in (1,\infty)$, the following holds:*

$$\|\boldsymbol{\mu}_s\|^2 \geq C_s \sqrt{\log \frac{1}{\delta}} \frac{\sigma^2 \sqrt{d}}{N_{\min}} \tag{8}$$

$$\|\boldsymbol{\mu}_c\|^2 \geq C_c \sigma^2 \sqrt{\log \frac{1}{\delta}} \max\left\{ Q^{-1}(\epsilon) \sqrt{\frac{d}{N_{\min}}}, \frac{\sqrt{d}}{N_{\min}}, \frac{\|\boldsymbol{\mu}_s\|^2}{N_{\min}\|\boldsymbol{\mu}_c\|^2} \right\}. \tag{9}$$

Then, with probability at least $1 - \delta$ over the choice of the training data, the robust error of the model returned by Algorithm 1 does not exceed ϵ.

Proof sketch. First, recall that the error of $f_{\mathbf{v}} = v_1 \cdot \mathbf{w}_1 + v_2 \cdot \mathbf{w}_2$ under $\mathbb{P}_\theta$ is

$$Q\left(\frac{\langle v_1 \cdot \mathbf{w}_1 + v_2 \cdot \mathbf{w}_2, \boldsymbol{\mu}_c\rangle}{\sigma \|v_1 \cdot \mathbf{w}_1 + v_2 \cdot \mathbf{w}_2\|} \left(1 + \theta \frac{\langle v_1 \cdot \mathbf{w}_1 + v_2 \cdot \mathbf{w}_2, \boldsymbol{\mu}_s\rangle}{\langle v_1 \cdot \mathbf{w}_1 + v_2 \cdot \mathbf{w}_2, \boldsymbol{\mu}_c\rangle} \right) \right).$$

Using the fact that $\mathbf{w}_e = N_e^{-1} \sum_{i\in S_e^{\text{trn}}} \mathbf{x}_i y_i = \boldsymbol{\mu}_c + \theta_e \boldsymbol{\mu}_s + \bar{\mathbf{n}}_e$ when $\bar{\mathbf{n}}_e$ is the mean of noise vectors drawn in Equation (1) for the data used in the first phase of learning, we can re-write these

[4] Intuitively, if $|\theta_1 - \theta_2| = 0$ then the two training environments are indistinguishable and we cannot hope to identify that the correlation induced by $\boldsymbol{\mu}_s$ is spurious. Otherwise, we expect $|\theta_1 - \theta_2|$ to have a quantifiable effect on our ability to generalize robustly. For simplicity of this exposition we assume that the gap is bounded away from zero; the full result in the Appendix is stated in terms of $|\theta_1 - \theta_2|$.

[5] This assumption makes sure we have some positive labels in each environment.

inner products as $\langle v_1 \cdot \mathbf{w}_1 + v_2 \cdot \mathbf{w}_2, \boldsymbol{\mu}_s \rangle = (v_1\theta_1 + v_2\theta_2)\|\boldsymbol{\mu}_s\|^2 + \langle \boldsymbol{\mu}_s, v_1\bar{\mathbf{n}}_1 + v_2\bar{\mathbf{n}}_2 \rangle$ and $\langle v_1 \cdot \mathbf{w}_1 + v_2 \cdot \mathbf{w}_2, \boldsymbol{\mu}_c \rangle = (v_1 + v_2)\|\boldsymbol{\mu}_c\|^2 + \langle \boldsymbol{\mu}_c, v_1\bar{\mathbf{n}}_1 + v_2\bar{\mathbf{n}}_2 \rangle$. This suggests that to obtain the desired bound on the robust error of the classifier returned by Algorithm 1, we must upper bound the ratio

$$\frac{(v_1^\star\theta_1 + v_2^\star\theta_2)\|\boldsymbol{\mu}_s\|^2 + \langle \boldsymbol{\mu}_s, v_1^\star\bar{\mathbf{n}}_1 + v_2^\star\bar{\mathbf{n}}_2 \rangle}{(v_1^\star + v_2^\star)\|\boldsymbol{\mu}_c\|^2 + \langle \boldsymbol{\mu}_c, v_1^\star\bar{\mathbf{n}}_1 + v_2^\star\bar{\mathbf{n}}_2 \rangle},$$

when $v_1^\star$ and $v_2^\star$ are the solutions to the optimization problem in Stage 2 of Algorithm 1. The terms involving inner-products with the noise terms are zero-mean and can be bounded using standard Gaussian concentration arguments. Therefore, the main effort of the proof is upper bounding

$$\frac{v_1^\star\theta_1 + v_2^\star\theta_2}{v_1^\star + v_2^\star} \cdot \frac{\|\boldsymbol{\mu}_s\|^2}{\|\boldsymbol{\mu}_c\|^2}.$$

To this end, we leverage the EOpp constraint. The population version of this constraint (corresponding to infintie N_1 and N_2) implies that $v_1^\star\theta_1 + v_2^\star\theta_2 = 0$. For finite sample sizes, we use standard Gaussian concentration and the Hanson-Wright inequality to show that the empirical EOpp constraint implies that $|v_1^\star\theta_1 + v_2^\star\theta_2|$ goes to zero as the sample sizes increases. Furthermore, we argue that $|v_1^\star + v_2^\star| \geq |\theta_1 - \theta_2|/2$, implying that—for appropriately large sample sizes—the above ratio indeed goes to zero. □

5 Empirical Validation

The empirical observations that motivated this work can be found across the literature. We thus focus our simulations on validating the theoretical results in our simplified model. We also evaluate Algorithm 1 on the Waterbirds dataset, where the goal is not to show state-of-the-art results, but rather to observe whether our formal claims hold beyond the Linear Two Environment Problem.

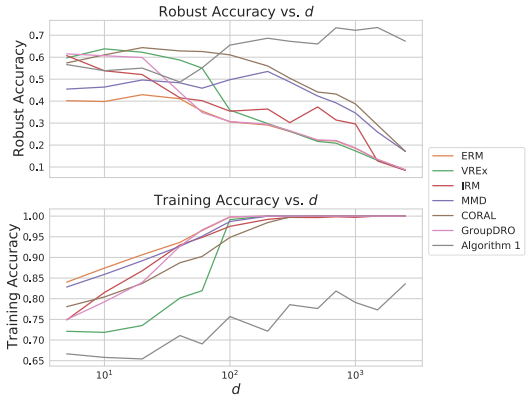

Figure 1: Numerical validation of our theoretical claims. Invariance inducing methods improve robust accuracy compared to ERM in low values of d, but their ability to do so is diminished as d grows (top plot) and they enter the interpolation regime, as seen on the bottom plot for $d > 10^2$. Algorithm 1 learns robust predictors as d grows and does not interpolate.

5.1 Simluations

Setup. Our simulation generates data as described in Theorem 1 with two environments where $\theta_1 = 1, \theta_2 = 0$. We further fix $r_c = 1$ and $r_c = 2$, while $N_1 = 800$ and $N_2 = 100$. We then take growing values of d, while adjusting σ so that $(r_c/\sigma)^2 \propto \sqrt{d/N}$.[6] For each value of d we train linear models with IRMv1 (Arjovsky et al., 2019), VREx (Krueger et al., 2020), MMD (Li et al., 2018), CORAL (Sun & Saenko, 2016), GroupDRO (Sagawa et al., 2019), implemented in the Domainbed package (Gulrajani & Lopez-Paz, 2020). We also train a classifier with the logistic loss to minimize empirical error (ERM), and apply Algorithm 1 where the "fine-tuning" stage trains a linear model over the two-dimensional representation using the VREx penalty to induce invariance. We repeat this for 15 random seeds to set μ_c, μ_s and to draw the training set.

Evaluation and results. We compare the robust accuracy and the train set accuracy of the learned classifiers as d grows. First, we observe that all methods except for Algorithm 1 attain perfect accuracy for large enough d, i.e. they interpolate. We further note that while invariance inducing methods give a desirable effect in low dimensions (the non-interpolating regime) – significantly improving the robust error over ERM – they become aligned with ERM in terms of robust accuracy as they go deeper into the interpolation regime (indeed, IRM essentially coincides with ERM for

[6]This is to keep our parameters within the regime where benign overfitting occurs.

larger d). This is an expected outcome considering our findings in section 3, as we set here N_1 to be considerably larger than N_2.

5.2 WATERBIRDS DATASET

We evaluate Algorithm 1 on the Waterbirds dataset (Sagawa et al., 2019), which has been previously used to evaluate the fairness and robustness of deep learning models.

Setup. Waterbirds is a synthetically created dataset containing images of water- and land-birds overlaid on water and land background. Most of the waterbirds (landbirds) appear in water (land) backgrounds, with a smaller minority of waterbirds (landbirds) appearing on land (water) backgrounds. The dataset is split into training, validation and test sets with 4795, 1199 and 5794 images in each set, respectively. We follow previous work (Sagawa et al., 2020; Veldanda et al., 2022) in defining a binary task in which waterbirds is the positive class and landbirds are the negative class, and using the following random features setup: for every image, a fixed pre-trained ResNet-18 model is used to extract a d_{rep}-dimensional feature vector $\mathbf{x}'$ ($d_{\text{rep}} = 512$). This feature vector is then converted into an d-dimensional feature vector $\mathbf{x} = \text{ReLU}(U\mathbf{x}')$, where $U \in \mathbb{R}^{d \times d_{\text{rep}}}$ is a random matrix with Gaussian entries. Finally, a logistic regression classifier is trained on $\mathbf{x}$. The extent of over-parameterization in this setup is controlled by varying d, the dimensionality of $\mathbf{x}$. In our experiments we vary d from 50 to 2500, with interpolation empirically observed at $d = 1000$ (which we refer to as the interpolation threshold).

Fairness. We use the image background type (water or land) as the sensitive feature, denoted A, and consider the fairness desiderata of Equal Opportunity Hardt et al. (2016), i.e., the false negative rate (FNR) should be similar for both groups. Towards this, we use the MinDiff penalty term (Prost et al., 2019). It uses the maximum mean discrepancy (MMD) distance between the model's output for the two sensitive groups when $Y = 1$ as a differentiable proxy to the FNR gap:

$$\mathcal{L}_M(\mathbf{w}) = \text{MMD}\left(\langle \mathbf{w}, X\rangle | A = 0, Y = 1;\ \langle \mathbf{w}, X\rangle | A = 1, Y = 1\right).$$

Evaluation. We compare the following methods: **(1) Baseline**: Learning a linear classifier $\mathbf{w}$ by minimizing $\mathcal{L}_p + \lambda \cdot \mathcal{L}_M$, where $\mathcal{L}_p$ is the standard binary cross entropy loss and $\mathcal{L}_M$ is the MinDiff penalty; **(2) Algorithm 1**: In the first stage, we learn group-specific linear classifiers $\mathbf{w}_0, \mathbf{w}_1$ by minimizing $\mathcal{L}_p$ on the examples from $A = 0$ and $A = 1$, respectively. In the second stage we learn $\boldsymbol{v} \in \mathbb{R}^2$ by minimizing $\mathcal{L}_p + \lambda \cdot \mathcal{L}_M$ on examples the entire dataset, where the new representation of the data is $\tilde{X} = [\langle w_1, X\rangle, \langle w_2, X\rangle] \in \mathbb{R}^2$.[7]

For all the experiments we use the Adam optimizer, a batch size of 128 and a learning rate schedule with initial rate of 0.01 and a decay factor of 10 for every 10,000 gradient steps. Every experiment is repeated 25 times and results are reported over all runs. For the baseline model we train for a total of 30,000 gradient steps whereas for our two-phased algorithm we use 15,000 gradient steps for each model in Phase A and an additional 250 steps for Phase B.

Results. Our main objective is to understand the effect of the fairness penalty. Towards this, for each method we compare both the test error and the test FNR gap when using either $\lambda = 0$ (no regularization) or $\lambda = 5$. The results are summarized in Figure 2. We can see that for the baseline approach, the fairness penalty successfully reduces the FNR gap when the classifier is not interpolating. However, as our negative result predicts and as previously reported in Veldanda et al. (2022), the fairness penalty becomes ineffective in the interpolating regime ($d \geq 1000$). On the other hand, for our two-phased algorithm, the addition of the fairness penalty reduces does reduce the FNR gap with an average relative improvement of 20%); crucially, this improvement is independent of d.

[7] This is basically Algorithm 1 with the following minor modifications: (1) The $\mathbf{w}_e$'s are computed via ERM, rather than simply taken to be the mean estimators; (2) Since the FNR gap penalty is already computed w.r.t a small number of samples, we avoid splitting the data and use the entire training set for both phases; (3) we convert the constrained optimization problem into an unconstrained problem with a penalty term.

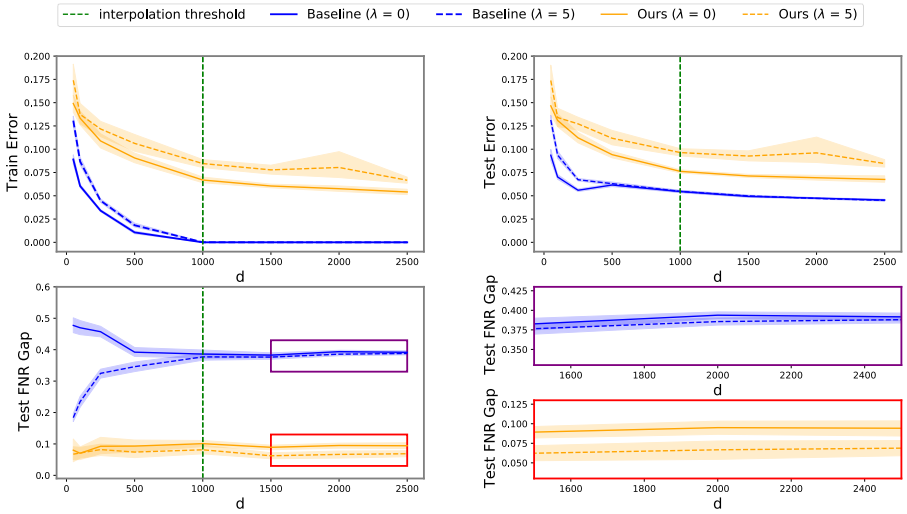

Figure 2: Results for the Waterbirds dataset (Sagawa et al., 2019). **Top row**: Train error (left) and test error (right). The train error is used to identify the interpolation threshold for the baseline method (approximately $d = 1000$). **Bottom row**: Comparing the FNR gap on the test set (left), with zoomed-in versions on the right. For the baseline approach, the fairness penalty successfully reduces the FNR gap when the classifier is not interpolating, but is ineffective in the interpolating regime ($d \geq 1000$). On the other hand, for our two-phased algorithm, the addition of the fairness penalty reduces the FNR gap in a way that is independent of d (average relative improvement 20%).

6 Discussion and Additional Related Work

In terms of formal results, most of the guarantees about invariant learning algorithms rely on the assumption that infinite training data is available (Arjovsky et al., 2019; Wald et al., 2021; Veitch et al., 2021; Puli et al., 2021; Rosenfeld et al., 2020). Some exceptions are the works of Ahuja et al. (2021) and Parulekar et al. (2022) that characterize the sample complexity of methods that learn invariant classifiers, yet they do not analyze the overparameterized cases we are concerned with. Negative results about learning overparameterized robust classifiers have been shown for methods based on importance weighting (Zhai et al., 2022), and negative results on learning with group-robust classifiers have been shown for max-margin classifiers (Sagawa et al., 2020). Our result is thus more general and applies to any learning algorithm that separates the data with arbitrarily small margins, instead of focusing on max-margin classifiers or specific algorithms. While we focus on the linear case, we believe it is instructive, as any reasonable method is expected to succeed in that case. Nonetheless, we believe our results can be extended to non-linear margins, and we leave this to future work.

One take-away from our result is that while low training loss is not something to avoid, overfitting to the point of interpolation creates a significant difficulty. This means one cannot assume a typical deep learning model with an added invariance penalty will indeed achieve any form of invariance; this fact also motivates using held-out data for imposing invariance, as in our Algorithm 1 as well as several other two-stage approaches mentioned above.

While our focus in this work was on theory underlying a wide array of algorithms, there are many closely related topics that we did not touch upon. For instance, an empirical comparison of two-stage methods along with other methods that avoid interpolation, e.g. by subsampling data (Idrissi et al., 2022; Chatterji et al., 2022). We also note that our focus in this paper was not on types of invariance that are satisfiable by using clever data augmentation techniques (e.g. invariance to image translation), or the design of special architectures (e.g. Cohen & Welling (2016); Lee et al. (2019); Maron et al. (2019)). These methods cleverly incorporate a-priori known invariances, and their empirical success when applied to large models may suggest that there are lessons to be learned for the type of invariant learning considered in our paper. These connections seem like an exciting avenue for future research.

REFERENCES

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pp. 145–155. PMLR, 2020.

Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R. Varshney. Empirical or invariant risk minimization? a sample complexity perspective. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jrA5GAccy_.

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

Keith Ball et al. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31 (1-58):26, 1997.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Ajay Byanjankar, Markku Heikkilä, and Jozsef Mezei. Predicting credit risk in peer-to-peer lending: A neural network approach. In *2015 IEEE symposium series on computational intelligence*, pp. 719–725. IEEE, 2015.

Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 872–881. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/byrd19a.html.

Yuan Cao, Quanquan Gu, and Misha Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=ChWy1anEuow.

Niladri S. Chatterji, Saminul Haque, and Tatsunori Hashimoto. Undersampling is a minimax optimal robustness intervention in nonparametric classification, 2022. URL https://arxiv.org/abs/2205.13094.

Valeriia Cherepanova, Vedant Nanda, Micah Goldblum, John P Dickerson, and Tom Goldstein. Technical challenges for training fair neural networks. *arXiv preprint arXiv:2102.06764*, 2021.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2189–2200. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/creager21a.html.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Yana Dranker, He He, and Yonatan Belinkov. Irm—when it works and when it doesn't: A test case of natural language inference. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18212–18224. Curran Associates, Inc., 2021.

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

Lin Lawrence Guo, Stephen R. Pfohl, Jason Fries, Alistair E. W. Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific Reports*, 12(1):2726, 2022.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

Jivat Neet Kaur, Emre Kiciman, and Amit Sharma. Modeling the data-generating process is necessary for out-of-distribution generalization. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022. URL https://openreview.net/forum?id=KfB7QnuseT9.

Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34, 2021.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations, 2022. URL https://arxiv.org/abs/2204.02937.

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.

Ananya Kumar, Tengyu Ma, Percy Liang, and Aditi Raghunathan. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift. In James Cussens and Kun Zhang (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1041–1051. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/kumar22a.html.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

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.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally-motivated shortcut removal using auxiliary labels. *arXiv preprint arXiv:2105.06422*, 2021.

Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the universality of invariant networks. In *International conference on machine learning*, pp. 4363–4371. PMLR, 2019.

Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Overparameterisation and worst-case generalisation: friend or foe? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jphnJNOwe36.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

Advait Parulekar, Karthikeyan Shanmugam, and Sanjay Shakkottai. Pac generalization via invariant representations, 2022. URL https://arxiv.org/abs/2205.15196.

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Flavien Prost, Hai Qian, Qiuwen Chen, Ed H Chi, Jilin Chen, and Alex Beutel. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*, 2019.

Aahlad Manas Puli, Lily H Zhang, Eric Karl Oermann, and Rajesh Ranganath. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. In *International Conference on Learning Representations*, 2021.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18347–18377. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/rame22a.html.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization, 2022. URL https://arxiv.org/abs/2202.06856.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8346–8356. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sagawa20a.html.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Ohad Shamir. The implicit bias of benign overfitting. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 448–478. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/shamir22a.html.

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.

Adarsh Subbaswamy, Bryant Chen, and Suchi Saria. A unifying causal framework for analyzing dataset shift-stable learning algorithms. *Journal of Causal Inference*, 10(1):64–89, 2022.

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

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=BdKxQp0iBi8.

Akshaj Kumar Veldanda, Ivan Brugere, Jiahao Chen, Sanghamitra Dutta, Alan Mishler, and Siddharth Garg. Fairness via in-processing in the over-parameterized regime: A cautionary tale. *arXiv preprint arXiv:2206.14853*, 2022.

Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, pp. 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *arXiv preprint arXiv:2102.10395*, 2021.

Ke Wang and Christos Thrampoulidis. Benign overfitting in binary classification of gaussian mixtures. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4030–4034. IEEE, 2021.

Ke Alexander Wang, Niladri Shekhar Chatterji, Saminul Haque, and Tatsunori Hashimoto. Is importance weighting incompatible with interpolating classifiers? In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. URL https://openreview.net/forum?id=pEhpLxVsd03.

Robert Williamson and Aditya Menon. Fairness risk measures. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6786–6797. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/williamson19a.html.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. Understanding why generalized reweighting does not improve over erm, 2022.

Jianyu Zhang, David Lopez-Paz, and Leon Bottou. Rich feature construction for the optimization-generalization dilemma. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26397–26411. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhang22u.html.

Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. Sparse invariant risk minimization. In *International Conference on Machine Learning*, pp. 27222–27244. PMLR, 2022.