An Investigation of Why Overparameterization Exacerbates Spurious Correlations

Shiori Sagawa [*1] **Aditi Raghunathan** [*1] **Pang Wei Koh** [*1] **Percy Liang** [1]

Abstract

We study why overparameterization—increasing model size well beyond the point of zero training error—can hurt test error on minority groups despite improving average test error when there are spurious correlations in the data. Through simulations and experiments on two image datasets, we identify two key properties of the training data that drive this behavior: the proportions of majority versus minority groups, and the signal-to-noise ratio of the spurious correlations. We then analyze a linear setting and theoretically show how the inductive bias of models towards "memorizing" fewer examples can cause overparameterization to hurt. Our analysis leads to a counterintuitive approach of subsampling the majority group, which empirically achieves low minority error in the overparameterized regime, even though the standard approach of upweighting the minority fails. Overall, our results suggest a tension between using overparameterized models versus using all the training data for achieving low worst-group error.

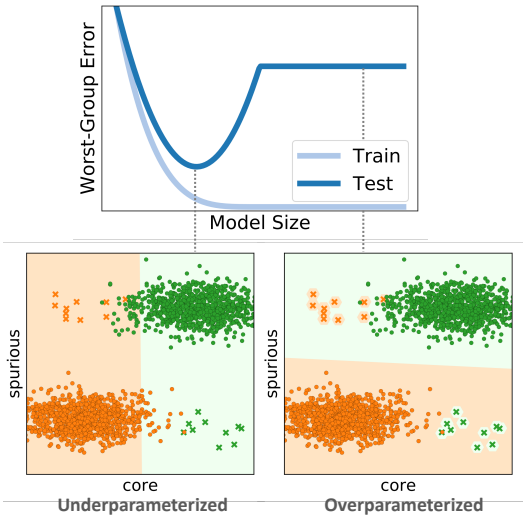

Figure 1. **Top**: Overparameterization *hurts* test error on the worst group when models are trained with the reweighted objective that upweights minority groups (Equation 3). Without reweighting, models have poor worst-group error regardless of model size (Appendix A.1). **Bottom**: Consider data points (x, y), where $x \in \mathbb{R}^2$ comprises a core feature x_{core} (x-axis) and a spurious feature x_{spu} (y-axis). The label y is highly correlated with x_{spu}, except on two minority groups (crosses). Underparameterized models use the core feature (left), but overparameterized models use the spurious feature and memorize the minority points (right).

1. Introduction

The typical goal in machine learning is to minimize the average error on a test set that is independent and identically distributed (i.i.d.) to the training set. A large body of prior work has shown that overparameterization—increasing model size beyond the point of zero training error—improves average test error in a variety of settings, both empirically (with neural networks, e.g., Nakkiran et al. (2019)) and theoretically (with linear and random projection models, e.g., Belkin et al. (2019); Mei & Montanari (2019)).

However, recent work has also demonstrated that models with low average error can still fail on particular groups of data points (Blodgett et al., 2016; Hashimoto et al., 2018; Buolamwini & Gebru, 2018). This problem of high worst-group error arises especially in the presence of spurious correlations, such as strong associations between label and background in image classification (McCoy et al., 2019; Sagawa et al., 2020). To mitigate this problem, common approaches reduce the worst-group training loss, e.g., through distributionally robust optimization (DRO) or simply upweighting the minority groups. Sagawa et al. (2020) showed these approaches improve worst-group error on strongly regularized neural networks but fail to help standard neural networks that can achieve zero training error, suggesting that increasing model capacity by reducing regularization—and perhaps by increasing overparameterization as well—can exacerbate spurious correlations.

[*]Equal contribution [1]Stanford University. Correspondence to: Shiori Sagawa <ssagawa@cs.stanford.edu>, Aditi Raghunathan <aditir@stanford.edu>, Pang Wei Koh <pangwei@cs.stanford.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

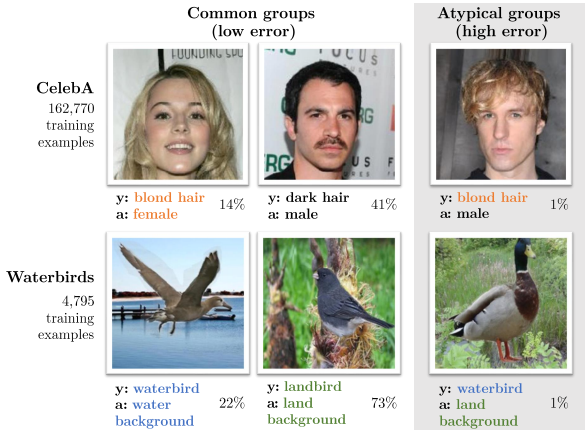

Figure 2. We consider two image datasets, CelebA and Waterbirds, where the label y is correlated with a spurious attribute a in a majority of the training data. The % beside each group shows its frequency in the training data. To measure how robust a model is to the spurious attribute, we divide the data into groups based on (y, a) and record the highest error incurred by a group. Figure adapted from Sagawa et al. (2020).

In this paper, we investigate why overparameterization exacerbates spurious correlations under the above approach of upweighting minority groups. We first confirm on two image datasets (Figure 2) that directly increasing overparameterization (i.e., increasing model size) indeed hurts worst-group error, leading to models that are highly inaccurate on the minority groups where the spurious correlation does not hold (Section 3). In contrast, their underparameterized counterparts obtain much better worst-group error, but do worse on average. We also confirm that models trained via empirical risk minimization (i.e., without upweighting the minority) have poor worst-group test error regardless of whether they are under- or overparameterized. Through simulations on a synthetic setting, we further identify two properties of the training data that modulate the effect of overparameterization: (i) the relative sizes of the majority versus minority groups, and (ii) how informative the spurious features are relative to the core features (Section 4).

Why does overparameterization exacerbate spurious correlations? Underparameterized models do not rely on spurious features because that would incur high training error on the (upweighted) minority groups where the spurious correlation does not hold. In contrast, overparameterized models can always obtain zero training error by memorizing training examples, and instead rely on their inductive bias to pick a solution—which features to use and which examples to memorize—out of all solutions with zero training error. Our results suggest an intuitive story of why overparameterization can hurt: because overparameterized models can have an inductive bias towards "memorizing" fewer examples (Figure 1). If (i) the majority groups are sufficiently large and (ii) the spurious features are more informative than the core features for these groups, then overparameterized models could choose to use the spurious features because it entails less memorization, and therefore suffer high worst-group test error. We test this intuition through simulations and formalize it in a theoretical analysis (Section 5).

Our analysis also leads to the counterintuitive result that on overparameterized models, subsampling the majority groups is much more effective at improving worst-group error than upweighting the minority groups. Indeed, an overparameterized model trained on a subset of <5% of the data performs similarly (on average and on the worst group) to an underparameterized model trained on all the data (Section 6). This suggests a possible tension between using overparameterized models and using all the data; average error benefits from both, but improving worst-group error seems to rely on using only one but not both.

2. Setup

Spurious correlation setup. We adopt the setting studied in Sagawa et al. (2020), where each example comprises the input features x, a label (core attribute) $y \in \mathcal{Y}$, and a spurious attribute $a \in \mathcal{A}$. Each example belongs to a group $g \in \mathcal{G} = \mathcal{Y} \times \mathcal{A}$, where $g = (y, a)$. Importantly, the spurious attribute a is correlated with the label y in the training set. We focus on the binary setting in which $\mathcal{Y} = \{1, -1\}$ and $\mathcal{A} = \{1, -1\}$.

Applications. We study two image classification tasks (Figure 2). In the first task, the label is spuriously correlated with demographics: specifically, we use the CelebA dataset (Liu et al., 2015) to classify hair color between the labels $\mathcal{Y} = \{\text{blonde, non-blonde}\}$, which are correlated with the gender $\mathcal{A} = \{\text{female, male}\}$. In the second task, the label is spuriously correlated with image background. We use the Waterbirds dataset (based on datasets from Wah et al. (2011); Zhou et al. (2017) and modified by Sagawa et al. (2020)) to classify between the labels $\mathcal{Y} = \{\text{waterbird, landbird}\}$, which are spuriously correlated with the image background $\mathcal{A} = \{\text{water background, land background}\}$. See Appendix A.5 for more dataset details.

Objectives and metrics. We evaluate a model w by its *worst-group* error,

$$\text{Err}_{\text{wg}}(w) := \max_{g \in \mathcal{G}} \mathbb{E}_{x,y|g} \left[\ell_{0-1}(w; (x, y)) \right], \qquad (1)$$

where ℓ_{0-1} is the 0-1 loss. In other words, we measure the error (% of examples that are incorrectly labeled) in each group, and then record the highest error across all groups. The standard approach to training models is empirical risk minimization (ERM): given a loss function ℓ, find the model w that minimizes the average training loss

$$\hat{\mathcal{R}}_{\text{ERM}}(w) = \hat{\mathbb{E}}_{(x,y,g)} \left[\ell(w; (x, y)) \right]. \qquad (2)$$

However, in line with Sagawa et al. (2020), we find that models trained via ERM have poor worst-group test error regardless of whether they are under- or overparameterized (Appendix A.1). To achieve low worst-group test error, prior work proposed modified objectives that focus on the worst-group loss, such as group distributionally robust optimization (group DRO) which directly optimizes for the worst-group training loss (Hu et al., 2018; Sagawa et al., 2020) or reweighting (Shimodaira, 2000; Byrd & Lipton, 2019). Sagawa et al. (2020) showed that both approaches can help worst-group loss, though group DRO is typically more effective. For simplicity, we focus on the well-studied reweighting approach, which optimizes

$$\hat{\mathcal{R}}_{\mathsf{reweight}}(w) = \hat{\mathbb{E}}_{(x,y,g)}\left[\frac{1}{\hat{p}_g}\ell(w;(x,y))\right], \qquad (3)$$

where $\hat{p}_g$ is the fraction of training examples in group g. The intuition behind reweighting is that it makes each group contribute the same weight to the training objective: that is, minority groups are upweighted, while majority groups are downweighted. Note that this approach requires the groups g to be specified at training time, though not at test time.

3. Overparameterization Hurts Worst-Group Error

Sagawa et al. (2020) observed that decreasing L_2 regularization hurts worst-group error. Though increasing overparameterization and reducing regularization can have different effects (Zhang et al., 2017; Mei & Montanari, 2019), this suggests that overparameterization might similarly exacerbate spurious correlations. Here, we show that directly increasing overparameterization (model size) indeed hurts worst-group error even though it improves average error.

Models. We study the CelebA and Waterbirds datasets described above. For CelebA, we train a ResNet10 model (He et al., 2016), varying model size by increasing the network width from 1 to 96, as in Nakkiran et al. (2019). For Waterbirds, we use logistic regression over random projections, as in Mei & Montanari (2019). Specifically, let $x \in \mathbb{R}^d$ denote the input features, which we obtain by passing the input image through a pre-trained, fixed ResNet-18 model. We train an unregularized logistic regression model over the feature representation $\mathrm{ReLU}(Wx) \in \mathbb{R}^m$, where $W \in \mathbb{R}^{m\times d}$ is a random matrix with each row sampled uniformly from the unit sphere $\mathbb{S}^{d-1}$. We vary model size by increasing the number of projections m from 1 to 10,000. We train each model by minimizing the reweighted objective (Equation (3)). For more details, see Appendix A.5.

Results. Overparameterization improves average test error across both datasets, in line with prior work (Belkin et al., 2019; Nakkiran et al., 2019) (Figure 3). However, in stark contrast, overparameterization *hurts* worst-group error: the best worst-group test error is achieved by an *underparameterized* model with non-zero training error. On CelebA, the smallest model (width 1) has 12.4% training error but comparatively low worst-group test error of 25.6%. As width increases, training error goes to zero but worst-group test error gets worse, reaching >60% for overparameterized models with zero training error. Similarly, on Waterbirds, an underparameterized model with 90 random features and training error of 17.7% obtains the best worst-group test error of 26.6%, while overparameterized models with zero training error yield worst-group test error of 42.4% at best.

In Appendix A.2, we also confirm that stronger regularization improves worst-group error but hurts average error in overparameterized models, while it has little effect on both worst-group and average error in underparameterized models. However, we focus on understanding the effect of overparameterization in the remainder of the paper.

Discussion. Why does overparameterization hurt worst-group test error? We make two observations. First, in the overparameterized regime, the smallest groups incur the highest test error (blonde males in CelebA and waterbirds on land background in Waterbirds), despite having zero training error. In other words, overparameterized models perfectly fit the minority points at training time, but seem to do so by using patterns that do not generalize. We informally refer to this behavior as "memorizing" the minority points.

Second, underparameterized models do obtain low worst-group error by learning patterns that generalize to both majority and minority groups. Therefore, overparameterized models should also be able to learn these patterns while attaining zero training error (e.g., by memorizing the training points that the underparameterized model cannot fit). Despite this, overparameterized models seem to learn patterns that generalize well on the majority but do not work on the minority (such as the spurious attributes a in Figure 2).

What makes overparameterized models memorize the minority instead of learning patterns that generalize well on both majority and minority groups? We study this question in the next two sections: in Section 4, we use simulations to understand properties of the data distribution that give rise to this trend, and in Section 5 we analyze a simplified linear setting and show how the inductive bias of models towards memorizing fewer points can lead to overparameterized models choosing to use spurious correlations.

4. Simulation Studies

The discussion in Section 3 suggests two properties of the training distribution that modulate the effect of overparameterization on worst-group error. Intuitively, overparameterized models should be more incentivized to use the spurious features and memorize the minority groups if (i) the propor-

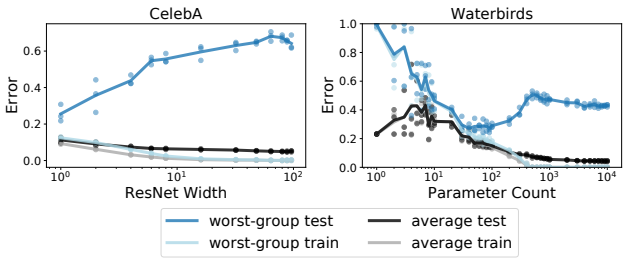

Figure 3. Increasing overparameterization (i.e., increasing model size) hurts the worst-group test error even though it improves the average test error. Here, we show results for models trained on the reweighted objective for CelebA (left) and Waterbirds (right).

tion of the majority group, p_{maj}, is higher, and (ii) the ratio of how informative the spurious features are relative to the core features, $r_{\text{s:c}}$, is higher. In this section, we use simulations to confirm these intuitions and probe how p_{maj} and $r_{\text{s:c}}$ affect worst-group error in overparameterized models.

4.1. Synthetic Experiment Setup

Data distribution. We construct a synthetic dataset that replicates the empirical trends in Section 3. As in Section 2, the label $y \in \{1, -1\}$ is spuriously correlated with a spurious attribute $a \in \{1, -1\}$. We divide our training data into four groups accordingly: two majority groups with $a = y$, each of size $n_{\text{maj}}/2$, and two minority groups with $a = -y$, each of size $n_{\text{min}}/2$. We define $n = n_{\text{maj}} + n_{\text{min}}$ as the total number of training points, and $p_{\text{maj}} = n_{\text{maj}}/n$ as the fraction of majority examples. The higher p_{maj} is, the more strongly a is correlated with y in the training data.

Each (y, a) group has its own distribution over input features $x = [x_{\text{core}}, x_{\text{spu}}] \in \mathbb{R}^{2d}$ comprising core features $x_{\text{core}} \in \mathbb{R}^d$ generated from the label/core attribute y, and spurious features $x_{\text{spu}} \in \mathbb{R}^d$ generated from the spurious attribute a:

$$x_{\text{core}} \mid y \sim \mathcal{N}(y\mathbf{1}, \sigma^2_{\text{core}} I_d)$$
$$x_{\text{spu}} \mid a \sim \mathcal{N}(a\mathbf{1}, \sigma^2_{\text{spu}} I_d). \quad (4)$$

The core and spurious features are both noisy and encode their respective attributes at different signal-to-noise ratios. We define the *spurious-core information ratio* (SCR) as $r_{\text{s:c}} = \sigma^2_{\text{core}}/\sigma^2_{\text{spu}}$. The higher the SCR, the more signal there is about the spurious attribute in the spurious features, relative to the signal about the label in the core features.

Compared to the image datasets we studied in Section 3, this synthetic dataset offers two key simplifications. First, the only differences between groups stem from their differences in (y, a), which isolates the effect of flipping the spurious attribute a. In contrast, in real datasets, groups can differ in other ways, e.g., more label noise in one group. Second, the relative difficulty of estimating y versus a is completely governed by changing σ^2_{core} and σ^2_{spu}. In contrast, real datasets have additional complications, e.g., estimating

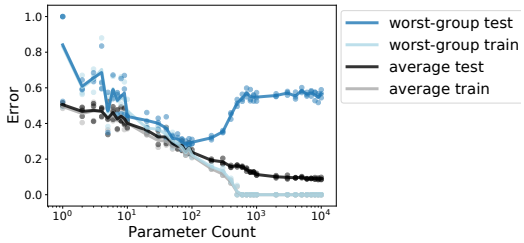

Figure 4. Overparameterization hurts worst-group test error but improves average test error on synthetic data, reproducing the trends we observe in real data.

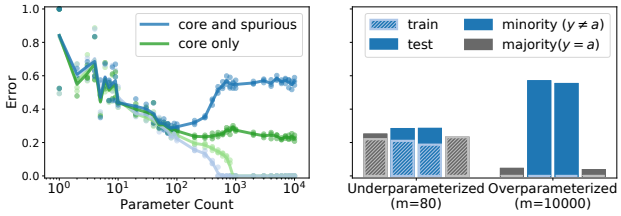

Figure 5. Overparameterized models have poor worst-group performance on the synthetic data because they rely on spurious features. **Left**: removing the spurious feature (green) eliminates the detrimental effect of overparameterization. **Right**: overparameterized models do well on the majority groups where the spurious features match the label, but poorly on the minority groups.

y might involve a more complex function of the input x than estimating a, and there might be an inductive bias towards learning a simpler model over a more complex one.

In all of the experiments below, we fix the total number of training points n to 3000, and set $d = 100$ (so each input x has $2d = 200$ dimensions). Unless otherwise specified, we set the majority fraction $p_{\text{maj}} = 0.9$ and the noise levels $\sigma^2_{\text{spu}} = 1$ and $\sigma^2_{\text{core}} = 100$ to encourage the model to use the spurious features over the core features.

Model. To avoid the complexities of optimizing neural networks, we follow the same random features setup we used for Waterbirds in Section 3: unregularized logistic regression using the reweighted objective on the random feature representation $\text{ReLU}(Wx) \in R^m$, where $W \in \mathbb{R}^{m \times d}$ is a random matrix (Mei & Montanari, 2019).

4.2. Observations on Synthetic Dataset

The synthetic dataset replicates the trends we observe on real datasets. Figure 4 shows how average and worst-group error change with the number of parameters/random projections m. This matches the trends we obtained on CelebA and Waterbirds in Section 3. The best worst-group test error of 28.5% is achieved by an underparameterized model, whereas highly overparameterized models achieve high worst-group test error that plateaus at around 55%. In contrast, the average test error is better for overparameterized models than for underparameterized models.

Overparameterized models use spurious features. Figure 5-Right shows that overparameterized models have high test error on minority groups ($a = -y$) despite zero training error, but perform very well on the majority groups ($a = y$). Since the only difference between the minority and majority groups in the synthetic dataset is the relative signs of the core and spurious attributes, this suggests overparameterized models are using spurious features and simply memorizing the minority groups to get zero training error, consistent with our discussion in Section 3. In contrast, the underparameterized model has low training and test errors across all groups, suggesting that it relies mainly on core features.

These results imply that the degradation in the worst-group test error is due to the spurious features. We confirm that overparameterization no longer hurts when we "remove" the spurious features by replacing them with noise centered around zero (i.e., we replace the mean of x_{spu} by 0). In this case, the best worst-group test error is now obtained by an overparameterized model, as shown in Figure 5-Left.

4.3. Distributional Properties

What properties of the training data make overparameterization hurt worst-group error? We study (i) p_{maj}, which controls the relative size of majority to minority groups, and (ii) $r_{\text{s:c}}$, the relative informativeness of spurious to core features. In the synthetic dataset, overparameterization hurts worst-group test error only when both are sufficiently high. In contrast, overparameterization helps average test error regardless; see Appendix A.3.

Effect of the majority fraction p_{maj}. We observe that increasing $p_{\text{maj}} = n_{\text{maj}}/n$, which controls the relative size of the majority versus minority groups, makes overparameterization hurt worst-group error more (Figure 6). When the groups are perfectly balanced with $p_{\text{maj}} = 0.5$, overparameterization no longer hurts the worst-group test error, with overparameterized models achieving better worst-group test error than all underparameterized models. This suggests that group imbalance can be a key factor inducing the detrimental effect of overparameterization.

Effect of the spurious-core information ratio $r_{\text{s:c}}$. Next, we characterize the effect of $r_{\text{s:c}} = \sigma^2_{\text{core}}/\sigma^2_{\text{spu}}$, which measures the relative informativeness of the spurious versus core features. A high $r_{\text{s:c}}$ means that the spurious features are more informative. We vary $r_{\text{s:c}}$ by changing σ^2_{spu} while keeping $\sigma^2_{\text{core}} = 100$ fixed, since this does not change the best possible worst-group test error (with a model that uses only the core features x_{core}). Figure 6 shows that the higher $r_{\text{s:c}}$ is, the more overparameterization hurts. As $r_{\text{s:c}}$ increases, the spurious features become more informative, and overparameterized models rely more on them than the core features; underparameterized models outperform overparameterized models only for sufficiently large $r_{\text{s:c}} \geq 1$. Note that increasing $r_{\text{s:c}}$ does not significantly affect the worst-group test error in the underparameterized regime, since the core features x_{core} are unaffected. In contrast, increasing the majority fraction p_{maj} hurts the worst-group test error in both underparameterized and overparameterized models.

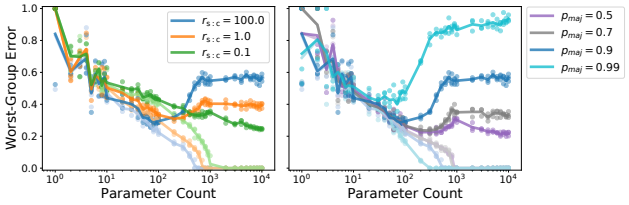

Figure 6. The higher the majority fraction p_{maj} and the spurious-core information ratio $r_{\text{s:c}}$, the more overparameterization hurts the worst-group test error. With sufficiently low p_{maj} and $r_{\text{s:c}}$, overparameterization switches to helping worst-group test error.

4.4. An Intuitive Story

We return to the question of what makes overparameterized models memorize the minority instead of learning patterns that generalize on both majority and minority groups. The simulation results above show that of all overparameterized models that achieve zero training error, the inductive bias of the model class and training algorithm favors models that use spurious features which generalize only for the majority groups, instead of learning to use core features that also generalize well on the minority groups.

What is the nature of this inductive bias? Consider a model that predicts the label y by returning its estimate of the spurious attribute a from x_{spu}, taking advantage of the fact that y and a are correlated in the training data. To get achieve zero training error, it will need to memorize the points in the minority group, e.g., by exploiting variations due to noise in the features x. On the other hand, consider a model that predicts y by returning a direct estimate of y based on the core features x_{core}. Because x_{core} provides a noisier estimate of y than x_{spu} does for a, this model will need to memorize all points for which x_{core} gives an inaccurate prediction of y due to noise. Since the estimators of the core and spurious attributes are equally easy to learn, the main difference between these two models is the number of examples to be memorized.

We therefore hypothesize that *the inductive bias favors memorizing as few points as possible*. This is consistent with the results above: the model uses x_{spu} and memorizes the minority points only when the fraction of minority points is small (high majority fraction p_{maj}). Similarly, the model uses x_{spu} over x_{core} to fit the majority points only when the spurious features are less noisy (high $r_{\text{s:c}}$) and therefore require less memorization to obtain zero training error than the core features. In the next section, we make this intuition

formal by analyzing a related but simpler linear setting.

5. Theoretical Analysis

In this section, we show how the inductive bias against memorization leads to overparameterization exacerbating spurious correlations. Our analysis explicates the effect of the inductive bias and the importance of the data parameters p_{maj} and $r_{\mathsf{s:c}}$ discussed in Section 4.

The synthetic setting discussed in Section 4 is difficult to analyze because of the non-linear random projections, so we introduce a linear *explicit-memorization* setting that allows us to precisely define the concept of memorization. For clarity, we refer to the previous synthetic setting in Section 4 as the *implicit-memorization* setting. In Appendix A.4, we show empirically that models in these two settings behave similarly in the overparameterized regime, though they differ in the underparameterized regime.

In the previous implicit-memorization setting, we varied model size and memorization capacity by varying the number of random projections of the input. In the new explicit-memorization setting, we instead use linear models that act directly on the input and introduce explicit "noise features" that can be used to memorize. We vary the memorization capacity by varying the number of explicit noise features.

5.1. Explicit-Memorization Setup

Training data. We consider input features $x = [x_{\mathsf{core}}, x_{\mathsf{spu}}, x_{\mathsf{noise}}]$, where the core feature $x_{\mathsf{core}} \in \mathbb{R}$ and the spurious feature $x_{\mathsf{spu}} \in \mathbb{R}$ are scalars. As in the implicit-memorization setup, they are generated based on the label and the spurious attribute, respectively:

$$x_{\mathsf{core}} \mid y \sim \mathcal{N}(y, \sigma^2_{\mathsf{core}}), \quad x_{\mathsf{spu}} \mid a \sim \mathcal{N}(a, \sigma^2_{\mathsf{spu}}).$$

The "noise" features $x_{\mathsf{noise}} \in \mathbb{R}^N$ are generated as

$$x_{\mathsf{noise}} \sim \mathcal{N}\left(0, \frac{\sigma^2_{\mathsf{noise}}}{N} I_N\right),$$

where $\sigma^2_{\mathsf{noise}}$ is a constant. The scaling by $1/N$ ensures that for large N, the norm of the noise vectors $\|x_{\mathsf{noise}}\|_2^2 \approx \sigma^2_{\mathsf{noise}}$ is approximately constant with high probability. Intuitively, when N is large, overparameterized models can use x_{noise} to fit a training point x without affecting its predictions on other points, thereby memorizing x. We formalize this notion of memorization later in Section 5.2.

As before, the training data is composed of four groups, each corresponding to a combination of the label $y \in \{-1, 1\}$ and the spurious attribute $a \in \{-1, 1\}$: two majority groups with $a = y$, each of size $n_{\mathsf{maj}}/2$, and two minority groups with $a = -y$, each of size $n_{\mathsf{min}}/2$. Combined, there are n training examples $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$.

Model. We study unregularized logistic regression on the input features $x \in \mathbb{R}^{N+2}$. As before, we consider the reweighted estimator $\hat{w}^{\mathsf{rw}}$. When the training data is linearly separable, the minimizer of the unregularized logistic loss on the training data is not well-defined. We therefore define $\hat{w}^{\mathsf{rw}}$ in terms of the sequence of L_2-regularized models $\hat{w}^{\mathsf{rw}}_\lambda$:

$$\hat{w}^{\mathsf{rw}}_\lambda \stackrel{\text{def}}{=} \arg\min_{w \in \mathbb{R}^{N+2}} \hat{\mathbb{E}}_{(x,y,g)} \left[\frac{1}{\hat{p}_g} \ell(w; (x, y))\right] + \frac{\lambda}{2}\|w\|_2^2,$$

where ℓ is the logistic loss and $\hat{p}_g$ is the fraction of training examples in group g. Since scaling a model does not affect its 0-1 error, we define $\hat{w}^{\mathsf{rw}}$ as the limit of this sequence, scaled to unit norm, as the regularization strength $\lambda \to 0^+$:

$$\hat{w}^{\mathsf{rw}} \stackrel{\text{def}}{=} \lim_{\lambda \to 0^+} \frac{\hat{w}^{\mathsf{rw}}_\lambda}{\|\hat{w}^{\mathsf{rw}}_\lambda\|_2}. \tag{5}$$

In the underparameterized regime, the training data is not linearly separable and we simply have $\hat{w}^{\mathsf{rw}} = \hat{w}^{\mathsf{rw}}_0 / \|\hat{w}^{\mathsf{rw}}_0\|_2$. In the overparameterized regime where $N \gg n$, the training data is linearly separable, and Rosset et al. (2004) showed that $\hat{w}^{\mathsf{rw}} = \hat{w}^{\mathsf{mm}}$, where $\hat{w}^{\mathsf{mm}}$ is the max-margin classifier

$$\hat{w}^{\mathsf{mm}} \stackrel{\text{def}}{=} \arg\max_{\|w\|_2 = 1} \min_i y^{(i)}(w \cdot x^{(i)}). \tag{6}$$

The equivalence $\hat{w}^{\mathsf{rw}} = \hat{w}^{\mathsf{mm}}$ holds regardless of the reweighting by $1/\hat{p}_g$: if we define the ERM estimator $\hat{w}^{\mathsf{erm}}$ analogously to (5) without the reweighting, it is also equal to $\hat{w}^{\mathsf{mm}}$. We will therefore analyze $\hat{w}^{\mathsf{mm}}$ in the overparameterized regime since it subsumes both $\hat{w}^{\mathsf{rw}}$ and $\hat{w}^{\mathsf{erm}}$.

We also note that if we use gradient descent to directly optimize the unregularized logistic regression objective (either reweighted or not), the resulting solution after scaling to unit norm also converges to $\hat{w}^{\mathsf{mm}}$ as the number of gradient steps goes to infinity (Soudry et al., 2018).

5.2. Analysis of Worst-Group Error

We now state our main analytical result: in the explicit-memorization setting, the worst-group test error of a sufficiently overparameterized model is greater than $1/2$ (worse than random) under certain settings of $\sigma^2_{\mathsf{spu}}, \sigma^2_{\mathsf{core}}, n_{\mathsf{maj}}, n_{\mathsf{min}}$. In contrast, underparameterized models attain reasonable worst-group error even under such a setting.

Theorem 1. *For any $p_{\mathsf{maj}} \geq \left(1 - \frac{1}{2001}\right)$, $\sigma^2_{\mathsf{core}} \geq 1$, $\sigma^2_{\mathsf{spu}} \leq \frac{1}{16 \log 100 n_{\mathsf{maj}}}$, $\sigma^2_{\mathsf{noise}} \leq \frac{n_{\mathsf{maj}}}{600^2}$ and $n_{\mathsf{min}} \geq 100$, there exists N_0 such that for all $N > N_0$ (overparameterized regime), with high probability over draws of the data,*

$$\mathit{Err}_{\mathsf{wg}}(\hat{w}^{\mathsf{mm}}) \geq 2/3, \tag{7}$$

where $\hat{w}^{\mathsf{mm}}$ is the max-margin classifier.

However, for $N = 0$ (underparameterized regime), with $p_{\mathsf{maj}} = \left(1 - \frac{1}{2001}\right)$, $\sigma^2_{\mathsf{core}} = 1$, and $\sigma^2_{\mathsf{spu}} = 0$, and in the

asymptotic regime with $n_{\mathsf{maj}}, n_{\mathsf{min}} \to \infty$, we have

$$Err_{\mathsf{wg}}(\hat{w}^{\mathsf{rw}}) < 1/4, \tag{8}$$

where $\hat{w}^{\mathsf{rw}}$ minimizes the reweighted logistic loss.

The result in the overparameterized regime applies to the max-margin classifier $\hat{w}^{\mathsf{mm}}$, which as discussed above subsumes both $\hat{w}^{\mathsf{rw}}$ and $\hat{w}^{\mathsf{erm}}$ when the data is linearly separable. The proof of Theorem 1 appears in Appendix B.

The conditions on σ^2_{spu} and σ^2_{core} in Theorem 1 above imply high spurious-core information ratio $r_{\mathsf{s:c}}$. Theorem 1 therefore provides a setting where high p_{maj} and high $r_{\mathsf{s:c}}$ provably make overparameterized models obtain high worst-group error, matching the trends we observed upon varying p_{maj} and $r_{\mathsf{s:c}}$ in the implicit-memorization setting (Figure 6). Furthermore, underparameterized models obtain reasonable worst-group error despite these conditions, mirroring the observations in earlier sections.

5.3. Overparameterization and Memorization

We now sketch the key ideas in the proof of Theorem 1 (full proof in Appendix B), focusing first on the overparameterized regime. We start by establishing an inductive bias towards learning the minimum-norm model that fits the training data. We then define memorization and show how the minimum-norm inductive bias translates into a bias against memorization. Finally, we illustrate how the bias against memorization leads to learning the spurious feature and suffering high worst-group error.

Minimum-norm inductive bias. Define a *separator* as any model that correctly classifies all of the training points (x, y) with margin $yw \cdot x \geq 1$. Then from standard duality arguments, $\hat{w}^{\mathsf{mm}}$ can be rewritten as $\hat{w}^{\mathsf{minnorm}}/\|\hat{w}^{\mathsf{minnorm}}\|$, the scaled version of the *minimum-norm separator* $\hat{w}^{\mathsf{minnorm}}$

$$\hat{w}^{\mathsf{minnorm}} \stackrel{\text{def}}{=} \underset{w \in \mathbb{R}^{N+2}}{\arg\min} \|w\|_2^2 \text{ s.t. } y^{(i)}(w \cdot x^{(i)}) \geq 1 \; \forall i. \tag{9}$$

Since scaling does not affect the 0-1 test error, it suffices to analyze $\hat{w}^{\mathsf{minnorm}}$. Equation (9) shows that out of the set of all separators (which all perfectly fit the training data), the inductive bias favors the separator with the minimum norm. We now discuss how this minimum-norm inductive bias favors less memorization.

Memorization. For convenience, we denote the three components of a model w as

$$w = \left[w_{\mathsf{core}}, w_{\mathsf{spu}}, w_{\mathsf{noise}}\right], \tag{10}$$

where $w_{\mathsf{core}} \in \mathbb{R}$, $w_{\mathsf{spu}} \in \mathbb{R}$, and $w_{\mathsf{noise}} \in \mathbb{R}^N$. By the representer theorem, we can decompose w_{noise} as follows:

$$w_{\mathsf{noise}} = \sum_i \alpha^{(i)} x^{(i)}_{\mathsf{noise}}. \tag{11}$$

In the overparameterized regime when $N \gg n$, a model can "memorize" a training point $x^{(i)}$ via w_{noise}, in particular by putting a large weight $\alpha^{(i)}$ in the direction of $x^{(i)}$ (Equation (11)):

Definition 1 (γ-memorization)**.** *A model w memorizes a point $x^{(i)}$ if $|\alpha^{(i)}| \geq \gamma^2/\sigma^2_{\mathsf{noise}}$ for some constant $\gamma \in \mathbb{R}$.*

Because the noise vectors of the training points (high-dimensional Gaussians) are nearly orthogonal for large N, the component $\alpha^{(i)} x^{(i)}_{\mathsf{noise}}$ affects the prediction on $x^{(i)}$, but not on any other training or test points.

This ability to memorize plays a crucial role in making overparameterized models obtain high worst-group error. Intuitively, the minimum-norm inductive bias favors less memorization in overparameterized models. Roughly speaking, models that memorize more have larger weights $|\alpha^{(i)}|$ on the noise vectors $x^{(i)}_{\mathsf{noise}}$. Since these noise vectors are nearly orthogonal and have similar norm, this translates into a larger norm $\|w_{\mathsf{noise}}\|_2^2$.

Comparing using x_{core} versus using x_{spu}. To illustrate how the inductive bias against memorization leads to high worst-group error, we consider two extreme sets of separators: (i) ones that use the spurious feature but not the core feature, denoted by $\mathcal{W}^{\mathsf{use-spu}}$ (ii) ones that use the core feature but not the spurious feature, denoted by $\mathcal{W}^{\mathsf{use-core}}$.

$$\begin{aligned}\mathcal{W}^{\mathsf{use-spu}} &\stackrel{\text{def}}{=} \{w \in \mathbb{R}^{N+2} : w \text{ is a separator}, w_{\mathsf{core}} = 0\} \\ \mathcal{W}^{\mathsf{use-core}} &\stackrel{\text{def}}{=} \{w \in \mathbb{R}^{N+2} : w \text{ is a separator}, w_{\mathsf{spu}} = 0\}.\end{aligned} \tag{12}$$

In scenario (i), using the spurious feature x_{spu} alone allows models to fit the majority groups very well. Thus, models that use x_{spu} only need to memorize the minority points. In Proposition 1, we construct a separator $w^{\mathsf{use-spu}} \in \mathcal{W}^{\mathsf{use-spu}}$ and show that its norm *only* scales with the number of minority points n_{min}.

Conversely, in scenario (ii), using the core feature x_{core} alone allows models to fit all groups equally well. However, when $r_{\mathsf{s:c}}$ is high, x_{core} is noisier than x_{spu}, so models that use x_{core} still need to memorize a constant fraction of *all* the training points. In Proposition 2, we show that norms of all separators $w^{\mathsf{use-core}} \in \mathcal{W}^{\mathsf{use-core}}$ are lower bounded by a quantity linear in the total number of training points n.

When the majority fraction p_{maj} is sufficiently large such that $n_{\mathsf{min}} \ll n$, the separator $w^{\mathsf{use-spu}}$ that uses x_{spu} will have a lower norm than any separator $w^{\mathsf{use-core}} \in \mathcal{W}^{\mathsf{use-core}}$ that uses x_{core}. Since the inductive bias favors the minimum-norm separator, it prefers a separator $w^{\mathsf{use-spu}}$ that memorizes the minority points and suffers high worst-group error over any $w^{\mathsf{use-core}} \in \mathcal{W}^{\mathsf{use-core}}$.

Proposition 1 (Norm of models using the spurious feature)**.** *When $\sigma^2_{\mathsf{core}}, \sigma^2_{\mathsf{spu}}$ satisfy the conditions in Theorem 1, there*

exists N_0 such that for all $N > N_0$, with high probability, there exists a separator $w^{\mathsf{use-spu}} \in \mathcal{W}^{\mathsf{use-spu}}$ such that

$$\|w^{\mathsf{use-spu}}\|_2^2 \leq \gamma_1^2 + \left(\frac{\gamma_2 n_{\mathsf{min}}}{\sigma^2_{\mathsf{noise}}}\right),$$

for some constants $\gamma_1, \gamma_2 > 0$.

Proof sketch. To simplify exposition in this sketch, suppose that the noise vectors $x^{(i)}_{\mathsf{noise}}$ are orthogonal and have constant norm $\|x^{(i)}_{\mathsf{noise}}\|_2^2 = \sigma^2_{\mathsf{noise}}$. We construct a separator $w^{\mathsf{use-spu}} \in \mathcal{W}^{\mathsf{use-spu}}$ that does not use the core feature x_{core} as follows. Set $w^{\mathsf{use-spu}}_{\mathsf{spu}} = \gamma_1$ for some large enough constant $\gamma_1 > 0$. This is sufficient to satisfy the margin condition on the majority points: since σ^2_{spu} is very small, w.h.p. all majority training points satisfy $y^{(i)}(x^{(i)}_{\mathsf{spu}}\gamma_1) \geq 1$.

However, for the minority training points, the spurious attribute a does not match the label y, and in order to satisfy the margin condition with a positive $w^{\mathsf{use-spu}}_{\mathsf{spu}}$, these n_{min} minority points have to be memorized. Since σ^2_{spu} is very small, the decrease in the margin due to $w^{\mathsf{use-spu}}_{\mathsf{spu}} = \gamma_1$ is at most $-\rho\gamma_1$ w.h.p. for some constant ρ that depends on σ^2_{spu}. To satisfy the margin condition, it thus suffices to set $\alpha^{(i)}_{\mathsf{use-spu}} = y^{(i)}(1+\rho\gamma_1)/\sigma^2_{\mathsf{noise}}$, and the bound on the norm follows. The full proof appears in Section B.2.6. □

Proposition 2 (Norm of models using the core feature)**.** *When $\sigma^2_{\mathsf{core}}, \sigma^2_{\mathsf{spu}}$ satisfy the conditions in Theorem 1 and $n_{\mathsf{min}} \geq 100$, there exists N_0 such that for all $N > N_0$, with high probability, all separators $w^{\mathsf{use-core}} \in \mathcal{W}^{\mathsf{use-core}}$ satisfy*

$$\|w^{\mathsf{use-core}}\|_2^2 \geq \frac{\gamma_3 n}{\sigma^2_{\mathsf{noise}}},$$

for some constant $\gamma_3 > 0$.

Proof sketch. Any model $w^{\mathsf{use-core}} \in \mathcal{W}^{\mathsf{use-core}}$ has $w^{\mathsf{use-core}}_{\mathsf{spu}} = 0$ by definition. We show that a constant fraction of training points have to be γ-memorized in order to satisfy the margin condition. We do so by first showing that the probability that a training point x satisfies the margin condition *without* being γ-memorized cannot be too large. For simplicity, suppose again that the noise vectors $x^{(i)}_{\mathsf{noise}}$ are orthogonal and have constant norm $\|x^{(i)}_{\mathsf{noise}}\|_2^2 = \sigma^2_{\mathsf{noise}}$. Then this probability is $\mathbb{P}\big(x_{\mathsf{core}} w^{\mathsf{use-core}}_{\mathsf{core}} \leq 1 - \gamma^2\big) \geq \Phi(-1/\sigma_{\mathsf{core}})$ for small γ, where Φ is the Gaussian CDF. Hence, in expectation, at least a constant fraction of points from the training distribution need to be memorized in order for $w^{\mathsf{use-core}}$ to satisfy the margin condition. With high probability, this is also true on the training set consisting of n points (via the DKW inequality) and the bound on the norm follows. The full proof appears in Section B.2.7. □

In the full proof of Theorem 1 in Appendix B, we generalize the above ideas to consider all separators in $\mathbb{R}^{N+2}$ instead of just the separators in $\mathcal{W}^{\mathsf{use-spu}} \bigcup \mathcal{W}^{\mathsf{use-core}}$. Note the importance of both $r_{\mathsf{s:c}}$ and p_{maj}: when $r_{\mathsf{s:c}}$ is high, models that use x_{spu} only need to memorize the minority groups (Proposition 1), and when p_{maj} is also high, these models end up memorizing fewer points than models that use x_{core} and have to memorize a constant fraction of the entire training set (Proposition 2).

6. Subsampling

Our results above highlight the role of the majority fraction p_{maj} in determining if overparameterization hurts worst-group test error. When p_{maj} is large, the inductive bias favors using spurious features because it entails memorizing only a relatively small number of minority points, while the alternative of using core features requires memorizing a large number of majority points. This suggests that reducing the memorization cost of using core features by directly removing some majority points could induce overparameterized models to obtain low worst-group error.

Here, we show that this approach of *subsampling* the majority group achieves good worst-group test error on the datasets studied above. Subsampling creates a new *group-balanced* dataset by randomly removing training points in all other groups to match the number of points from the smallest group (Japkowicz & Stephen, 2002; Haixiang et al., 2017; Buda et al., 2018). We then train a model to minimize the average loss on this subsampled dataset. For a precise description, see Appendix A.6.

Figure 7 shows that overparameterized models trained via subsampling (Equation 15) obtain low worst-group error on the CelebA, Waterbirds, and synthetic (implicit-memorization) datasets. Across all three datasets, training via subsampling makes increasing overparameterization help *both* average and worst-group test error. Moreover, overparameterized models trained on subsampled data are comparable to or better than the best models trained on the full dataset (i.e., underparameterized models trained with reweighting).

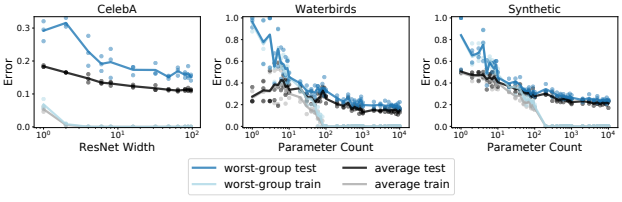

Figure 7. Overparameterization helps worst-group test error when training via subsampling, which involves creating a group-balanced dataset by reducing the number of majority points and minimizing average training loss on the new dataset.

Subsampling seems wasteful since it throws away a large fraction of the training data: we only use 3.4% of the full training data for CelebA, 4.6% for Waterbirds, and 10% for the synthetic dataset. However, the results above show that subsampling in overparameterized models matches or outperforms reweighting with underparameterized models. For example, on CelebA, an overparameterized model trained via subsampling obtains 11.1% average test and 15.1% worst-group test error, whereas an underparameterized model trained with reweighting obtains 11.3% average and 25.6% worst-group test error.

Subsampling vs. reweighting. Both subsampling and reweighting artificially balance the groups in the training data, and previous work on imbalanced datasets has concluded that reweighting is typically at least as effective as subsampling (Buda et al., 2018). However, we find a clear difference between subsampling and reweighting in the overparameterized regime: increasing overparameterization with reweighting increases worst-group error, while doing so with subsampling decreases worst-group error. The intuition developed in Sections 4 and 5 shed some light on this difference. Consider an overparameterized model: as in Section 5.1, reweighting does not change the learned model which is the max-margin classifier. However, subsampling reduces p_{maj}. Recall that the inductive bias favors spurious features when the alternative of using core features requires memorizing a large number of training points. By reducing p_{maj}, we reduce this memorization cost associated with core features, thereby inducing the model to use core features and achieve low worst-group test error.

7. Related Work

The effect of overparameterization. The effect of overparameterization on average test error has been widely studied. In what is commonly referred to as "double descent", increasing model size beyond zero training error decreases test error, despite conventional wisdom that overfitting should increase test error. This behavior has been observed empirically (Belkin et al., 2019; Opper, 1995; Advani & Saxe, 2017; Nakkiran et al., 2019) and shown analytically in high-dimensional regression (Hastie et al., 2019; Bartlett et al., 2019; Mei & Montanari, 2019). These works focus on average test error and are consistent with our findings there. However, our focus is on worst-group test error, particularly when the groups are defined based on spurious attributes, and in this paper we establish that worst-group test error can behave quite differently from average test error.

Increasing overparameterization can actually improve model robustness to some types of distributional shifts (Hendrycks et al., 2019; Hendrycks & Dietterich, 2019; Yang et al., 2020). In this light, our results show that the effect of overparameterization on model robustness can depend heavily on the dataset (e.g., properties like p_{maj} and $r_{\mathsf{s:c}}$), type of distributional shift, and training procedure.

Worst-group error. Prior work on improving worst-group error focused on the underparameterized regime, with methods based on weighting/sampling (Shimodaira, 2000; Japkowicz & Stephen, 2002; Buda et al., 2018; Cui et al., 2019), distributionally robust optimization (DRO) (Ben-Tal et al., 2013; Namkoong & Duchi, 2017; Oren et al., 2019), and fair algorithms (Dwork et al., 2012; Hardt et al., 2016; Kleinberg et al., 2017). Our focus is on the overparameterized, zero-training-error regime; here, previous methods based on reweighting and DRO are ineffective (Wen et al., 2014; Byrd & Lipton, 2019; Sagawa et al., 2020). As mentioned in Section 1, Sagawa et al. (2020) demonstrated that stronger L_2-regularization can improve worst-group error on neural networks (when coupled with reweighting or group DRO). Similarly Cao et al. (2019) show that data-dependent regularization can improve error on rare labels. While their work focuses on developing methods to improve worst-group error, our focus is on understanding the mechanisms by which overparameterization hurts worst-group error.

8. Discussion

Our work shows that overparameterization hurts worst-group error on real datasets that contain spurious correlations. We studied the implicit- and explicit-memorization settings to provide a potential story for why this might occur: there can be an inductive bias towards solutions that do not need to memorize as many training points, and this can favor models that exploit the spurious correlations.

However, our synthetic settings make several simplifying assumptions, e.g., they suppose that the model prefers the spurious feature because it is less noisy than the core feature. This assumption need not always apply, and different assumptions might also lead to overparameterization exacerbating spurious correlations. For example, there might exist a true classifier based on the core features which has high accuracy but which is relatively more complex (e.g., high parameter norm) and therefore not favored by the training procedure. Studying the effect of overparameterization in settings such as those is important future work.

We also observed that subsampling allows overparameterized models to achieve low average and worst-group test error, despite eliminating a large fraction of training examples. In contrast, when using the full training data, only underparameterized models attain low worst-group test error under our current training methods. These observations call for future work to develop methods that can exploit both the statistical information in the full training data as well as the expressivity of overparameterized models, so as to attain good worst-group and average test error.

Acknowledgements

We are grateful to Yair Carmon, John Duchi, Tatsunori Hashimoto, Ananya Kumar, Yiping Lu, Tengyu Ma, and Jacob Steinhardt for helpful discussions and suggestions. SS was supported by a Stanford Graduate Fellowship, AR was supported by a Google PhD Fellowship and Open Philanthropy Project AI Fellowship, and PWK was supported by the Facebook Fellowship Program.

Reproducibiltity

Code is available at https://github.com/ssagawa/overparam_spur_corr.
All code, data, and experiments are available on the Codalab platform at https://worksheets.codalab.org/worksheets/0x1db77e603a8d48c8abebd67fce39cf8b.

References

Advani, M. S. and Saxe, A. M. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *arXiv*, 2019.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Science*, 116(32), 2019.

Ben-Tal, A., den Hertog, D., Waegenaere, A. D., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59:341–357, 2013.

Blodgett, S. L., Green, L., and O'Connor, B. Demographic dialectal variation in social media: A case study of African-American English. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1119–1130, 2016.

Buda, M., Maki, A., and Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Byrd, J. and Lipton, Z. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning (ICML)*, pp. 872–881, 2019.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Cui, Y., Jia, M., Lin, T., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, 2019.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pp. 214–226, 2012.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

Hardt, M., Price, E., and Srebo, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3315–3323, 2016.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, 2018.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning (ICML)*, 2018.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pp. 448–456, 2015.

Japkowicz, N. and Stephen, S. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science (ITCS)*, 2017.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

McCoy, R. T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Association for Computational Linguistics (ACL)*, 2019.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

Namkoong, H. and Duchi, J. Variance regularization with convex objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Opper, M. Statistical mechanics of learning: Generalization. *The Handbook of Brain Theory and Neural Networks,*, pp. 922–925, 1995.

Oren, Y., Sagawa, S., Hashimoto, T., and Liang, P. Distributionally robust language modeling. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Rosset, S., Zhu, J., and Hastie, T. J. Margin maximizing loss functions. In *Advances in neural information processing systems*, pp. 1237–1244, 2004.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244, 2000.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research (JMLR)*, 19(1):2822–2878, 2018.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 15(1):1929–1958, 2014.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD Birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.

Wen, J., Yu, C., and Greiner, R. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning (ICML)*, pp. 631–639, 2014.

Yang, Z., Yu, Y., You, C., Steinhardt, J., and Ma, Y. Rethinking bias-variance trade-off for generalization of neural networks. *arXiv preprint arXiv:2002.11328*, 2020.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.