WHY ADVERSARIAL TRAINING CAN HURT ROBUST ACCURACY

Jacob Clarysse[1], Julia Hörrmann[2], Fanny Yang[1]
1. Department of Computer Science, ETH Zürich
2. Department of Mathematics, ETH Zürich
{jacob.clarysse;fan.yang}@inf.ethz.ch;
{julia.hoerrmann}@stat.math.ethz.ch

ABSTRACT

Machine learning classifiers with high test accuracy often perform poorly under adversarial perturbations. It is commonly believed that adversarial training alleviates this issue. In this paper, we demonstrate that, surprisingly, the opposite can be true for a natural class of perceptible perturbations — even though adversarial training helps when enough data is available, it may in fact hurt robust generalization in the small sample size regime. We first prove this phenomenon for a high-dimensional linear classification setting with noiseless observations. Using intuitive insights from the proof, we could find perturbations on standard image datasets for which this behavior persists. Specifically, it occurs for perceptible perturbations that effectively reduce class information such as object occlusions or corruptions.

1 INTRODUCTION

Today's best-performing classifiers are vulnerable to adversarial attacks Goodfellow et al. (2015); Szegedy et al. (2014) and exhibit high *robust error*: for many inputs, their predictions change under adversarial perturbations, even though the true class stays the same. Such content-preserving (Gilmer et al., 2018), consistent (Raghunathan et al., 2020) attacks can be either perceptible or imperceptible. For image datasets, most work to date studies imperceptible attacks that are based on perturbations with limited strength or *attack budget*. These include bounded ℓ_p-norm perturbations (Goodfellow et al., 2015; Madry et al., 2018; Moosavi-Dezfooli et al., 2016), small transformations using image processing techniques (Ghiasi et al., 2019; Zhao et al., 2020; Laidlaw et al., 2021; Luo et al., 2018) or nearby samples on the data manifold (Lin et al., 2020; Zhou et al., 2020). Even though they do not visibly change the image by definition, imperceptible attacks can often successfully fool a learned classifier.

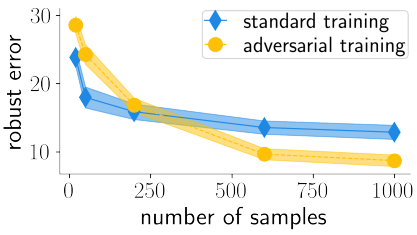


Figure 1: On the Waterbirds dataset attacked by the adversarial illumination attack, adversarial training (yellow) yields higher robust error than standard training (blue) when the sample size is small, even though it helps for large sample sizes and in a setting where the standard error of standard training is small. (see App. D for details).

On the other hand, perturbations that naturally occur and are physically realizable are commonly perceptible. Some perceptible perturbations specifically target the object to be recognized: these include occlusions (e.g. stickers placed on traffic signs (Eykholt et al., 2018) or masks of different sizes that cover important features of human faces (Wu et al., 2020)) or corruptions that are caused by the image capturing process (animals that move faster than the shutter speed or objects that are not well-lit, see Figure 2). Others transform the whole image and are not confined to the object itself, such as rotations, translations or corruptions Engstrom et al. (2019); Kang et al. (2019). In this paper, we refer to such perceptible attacks as *directed attacks*. In contrast to other attacks, they effectively reduce useful class information in the input for any model, without necessarily changing the true label — we say that they are *directed* and *consistent*, more formally defined in Section 2. For example, a stop sign with a small sticker could partially cover the text without losing its semantic meaning.

(a) Masks (b) Original (c) Lighting (d) Blur (e) Classification of perturbations

Figure 2: Examples of directed attacks on CIFAR10 and the Waterbirds dataset. In Figure 2a, we corrupt the image with a black mask of size 2×2 and in Figure 2c and 2d we change the lighting conditions (darkening) and apply motion blur on the bird in the image respectively. All perturbations reduce the information about the class in the images: they are the result of directed attacks. (e) Directed attacks are a subset of perceptible attacks.

Similarly, a flying bird captured with a long exposure time can induce motion blur in the final image without becoming unrecognizable to the observer.

In the literature so far, it is widely acknowledged that adversarial training with the same perturbation type and budget as during test time often achieves significantly lower robust error than standard training (Madry et al., 2018; Zhang et al., 2019; Bai et al., 2021).

In contrast, we show that adversarial training not only increases standard error (Zhang et al., 2019; Tsipras et al., 2019; Stutz et al., 2019; Raghunathan et al., 2020), but surprisingly, in the low sample regime,

adversarial training may even increase the robust error compared to standard training!

Figure 1 illustrates the main message of our paper on the Waterbirds dataset: Although adversarial training with directed attacks outperforms standard training when enough training samples are available, it is inferior when the sample size is small (but still large enough to obtain a small standard test error).

Our contributions are as follows:

- We prove that, almost surely, adversarially training a linear classifier on separable data yields a monotonically increasing robust error as the perturbation budget grows. We further establish high-probability non-asymptotic lower bounds on the robust error gap between adversarial and standard training.
- Our proof provides intuition for why this lower bound on the gap is particularly large for directed attacks in the low sample regime.
- We observe empirically for different directed attacks on real-world image datasets that this behavior persists: adversarial training for directed attacks hurts robust accuracy when the sample size is small.

2 Robust Classification

We first introduce our robust classification setting more formally by defining the notions of adversarial robustness, directed attacks and adversarial training used throughout the paper.

Adversarially robust classifiers For inputs $x \in \mathbb{R}^d$, we consider multi-class classifiers associated with parameterized functions $f_\theta : \mathbb{R}^d \to \mathbb{R}^K$ if $K > 2$ and $f_\theta : \mathbb{R}^d \to \mathbb{R}$ if $K = 2$, where K is the number of labels. For example, $f_\theta(x)$ could be a linear model (as in Section 3) or a neural network (as in Section 4). The output label predictions are obtained by $h(f_\theta(x)) = \text{sign}(f_\theta(x))$ for $K = 2$ and $h(f_\theta(x)) = \arg\max_{k \in \{1,..,K\}} f_\theta(x)_k$ for $K > 2$.

In order to convince practitioners to use machine learning models in the wild, it is key to demonstrate that they exhibit robustness. One kind of robustness is that they do not change prediction when the input is subject to consistent perturbations, which are small class-preserving perturbations. Mathematically speaking, for the underlying joint data distribution $\mathbb{P}$, the model should have a small ϵ_{te}*-robust error*, defined as

$$\text{Err}(\theta; \epsilon_{\text{te}}) := \mathbb{E}_{(x,y)\sim\mathbb{P}} \max_{x' \in T(x;\epsilon_{\text{te}})} \ell(f_\theta(x'), y), \tag{1}$$

where ℓ is 0 if the class determined by $h(f_\theta(x))$ is equal to y and 1 otherwise. Further, $T(x; \epsilon_{te})$ indicates a perturbation set around x of a certain transformation type with size ϵ_{test}. Note that the *(standard) error* $\mathbb{E}_{(x,y)\sim\mathbb{P}}\ell(f_\theta(x), y)$ of a classifier corresponds to $\text{Err}(\theta; 0)$ – the robust error evaluated at $\epsilon_{\text{te}} = 0$.

Directed attacks The inner maximization in Equation 1 is often called the adversarial *attack* of the input x for the model f_θ and the corresponding solution is referred to as the *adversarial example*. In this paper, we consider *directed attacks* that effectively reduce the information about the true classes, with image-based examples depicted in Figure 2. For linear classification, we analyze directed attacks in the form of additive perturbations that are constrained to the direction of the optimal decision boundary (see details in Section 3.1). In particular, note that the set of directed perturbations is restricted to directions attacking the Bayes optimal classifier.

Adversarial training A common approach to obtain classifiers with a good robust accuracy is to minimize the training objective $\mathcal{L}_{\epsilon_{\text{tr}}}$ with a surrogate robust classification loss L

$$\mathcal{L}_{\epsilon_{\text{tr}}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \max_{x_i' \in T(x_i;\epsilon_{\text{tr}})} L(f_\theta(x_i')y_i), \tag{2}$$

also called *adversarial training*. In practice, we often use the cross entropy loss $L(z) = \log(1 + e^{-z})$ and minimize the robust objective by using first order optimization methods such as (stochastic) gradient descent. SGD is also the algorithm that we focus on in both the theoretical and experimental sections. When the desired type of robustness is known in advance, it is standard practice to use the same perturbation set for training as for testing, i.e. $T(x; \epsilon_{\text{tr}}) = T(x; \epsilon_{\text{te}})$. For example, Madry et al. (2018) show that the robust error sharply increases for $\epsilon_{\text{tr}} < \epsilon_{\text{te}}$. In this paper, we demonstrate that for directed attacks in the small sample size regime, in fact, the opposite is true.

3 Theoretical Results

In this section, we prove for linear functions $f_\theta(x) = \theta^\top x$ that in the case of directed attacks, robust generalization deteriorates with increasing ϵ_{tr}. The proof, albeit in a simple setting, provides explanations for why adversarial training fails in the high-dimensional regime for such attacks.

3.1 Setting

We now introduce the precise linear setting used in our theoretical results.

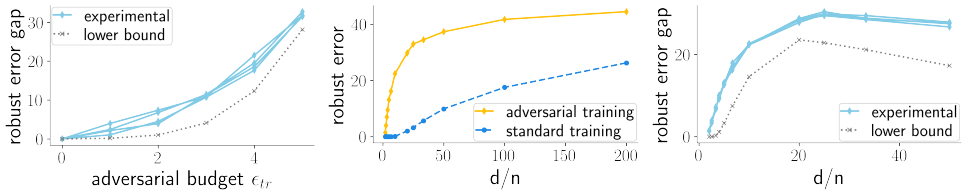
Data model We assume that the ground truth and hypothesis class are given by linear functions $f_\theta(x) = \theta^\top x$ and the sample size n is lower than the ambient dimension d minus one. The generative distribution $\mathbb{P}_r$ is similar to (Tsipras et al., 2019; Nagarajan & Kolter, 2019): The label $y \in \{+1, -1\}$ is drawn with equal probability and the covariate vector is sampled as $x = [y\frac{r}{2}, \tilde{x}]$ with the random vector $\tilde{x} \in \mathbb{R}^{d-1}$ drawn from a standard normal distribution, i.e. $\tilde{x} \sim \mathcal{N}(0, \sigma^2 I_{d-1})$. We would like to learn a classifier that has low robust error by using a dataset $D = (x_i, y_i)_{i=1}^n$ with n i.i.d. samples from $\mathbb{P}_r$. Intuitively, the separation distance r reflects the signal strength of the data distribution.

Notice that the distribution $\mathbb{P}_r$ is noiseless: for a given input x, the label $y = \text{sign}(x_{[1]})$ is deterministic. Further, the Bayes optimal linear classifier (also referred to as the *ground truth*) is parameterized by the first standard coordinate vector, $\theta^\star = e_1$.[1] By definition, the ground truth is robust against all perturbations that do not change the sign in the first coordinate of the sample, i.e. consistent perturbations, and hence so is the optimal robust classifier.

Directed attacks In this paper, we focus on consistent directed attacks that by definition efficiently concentrate their attack budget to reduce the class information. For our linear setting this information lies in the first entry. Hence, we can model such attacks by additive perturbations in the first dimension

$$T(x; \epsilon) = \{x' = x + \delta \mid \delta = \beta e_1 \text{ and } -\epsilon \leq \beta \leq \epsilon\}. \tag{3}$$

[1] Note that the result more generally holds for non-sparse models that are not axis aligned by way of a simple rotation $z = Ux$. In that case the distribution is characterized by $\theta^\star = u_1$, where u_1 is the first column vector of U, and a rotated Gaussian in the $d-1$ dimensions orthogonal to $\theta^\star$.



(a) Robust error increase with ϵ_{tr} (b) Standard-adversarial training (c) Effect of overparameterization

Figure 3: Experimental verification of Theorem 3.1. (a) We set $d = 1000$, $r = 12$ and $n = 50$. The robust error gap between standard and adversarial training as a function of the adversarial budget $\epsilon_{\text{tr}} = 5$ independent experiments (blue) and the lower bound given in Theorem 3.1 (gray). In (b) and (c), we set $d = 10000$ and vary the number of samples n. (b) The robust error of standard and adversarial training with $\epsilon_{\text{tr}} = 4.5$. (c) The error gap and the lower bound of Theorem 3.1. For more experimental details see Appendix C.

Note that this attack is always in the direction of the signal dimension, i.e. the Bayes optimal classifier or equivalently the ground truth. Furthermore, when $\epsilon < \frac{r}{2}$, it is a consistent directed attack. Observe how this is different from ℓ_p-attacks — an ℓ_p attack, depending on the model, may add a perturbation that only has a very small component in the signal direction.

Robust max-ℓ_2-margin classifier We study a classifier that is the solution of running gradient descent on the adversarial logistic loss. A long line of work (Soudry et al., 2018; Ji & Telgarsky, 2019; Chizat & Bach, 2020; Nacson et al., 2019; Liu et al., 2020) studies the implicit bias of (S)GD on the (standard) logistic loss and separable data. In particular, they show directional convergence to the max-margin solution. For the adversarial logistic loss and linear models in particular, (S)GD converges to the robust max-ℓ_2-margin solution (Li et al., 2020),

$$\widehat{\theta}^{\epsilon_{\text{tr}}} := \arg\max_{\|\theta\|_2 \le 1} \min_{i \in [n], x_i' \in T(x_i; \epsilon_{\text{tr}})} y_i \theta^\top x_i'. \tag{4}$$

Even though our result is proven for the max-ℓ_2-margin classifier, it can easily be extended to other interpolators.

3.2 Main results

We are now ready to characterize the ϵ_{te}-robust error as a function of ϵ_{tr}, the separation r, the dimension d and sample size n of the data. In the theorem statement we use the following quantities

$$\varphi_{\text{min}} = \frac{\sigma}{r/2 - \epsilon_{\text{te}}} \left(\sqrt{\frac{d-1}{n}} - \left(1 + \sqrt{\frac{2\log(2/\delta)}{n}} \right) \right)$$
$$\varphi_{\text{max}} = \frac{\sigma}{r/2 - \epsilon_{\text{te}}} \left(\sqrt{\frac{d-1}{n}} + \left(1 + \sqrt{\frac{2\log(2/\delta)}{n}} \right) \right)$$

that arise from concentration bounds for the singular values of the random data matrix. Further, let $\tilde{\epsilon} := \frac{r}{2} - \frac{\varphi_{\text{max}}}{\sqrt{2}}$ and denote by Φ the cumulative distribution function of a standard normal.

Theorem 3.1. *Assume $d - 1 > n$. For test samples from $\mathbb{P}_r$, perturbation set type T as in Equation 3 and any $0 \le \epsilon_{te} < \frac{r}{2}$, the following holds for the ϵ_{te}-robust error of the classifier (Equation 1) resulting from ϵ_{tr}-adversarial training:*

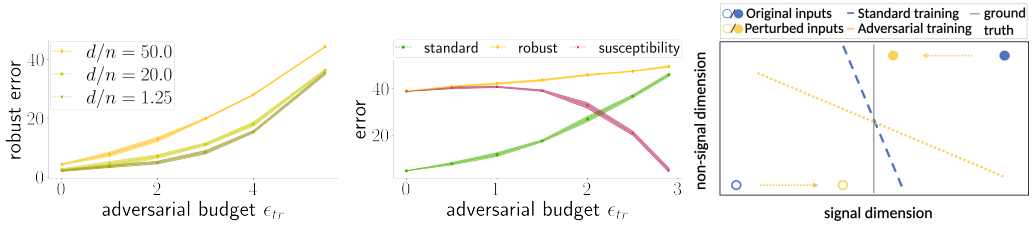
1. *The ϵ_{te}-robust error of the ϵ_{tr}-robust max-margin estimator reads*

$$Err(\widehat{\theta}^{\epsilon_{tr}}; \epsilon_{te}) = \Phi\left(-\frac{\left(\frac{r}{2} - \epsilon_{tr}\right)}{\tilde{\varphi}} \right) \tag{5}$$

   for a random quantity $\tilde{\varphi} > 0$ depending on σ, r, ϵ_{te} and is hence strictly increasing in the adversarial training budget ϵ_{tr}.

2. *With probability at least $1 - \delta$, we further have $\varphi_{min} \le \tilde{\varphi} \le \varphi_{max}$ and the following lower bound on the robust error increase by adversarially training with size ϵ_{tr}*

$$Err(\widehat{\theta}^{\epsilon_{tr}}; \epsilon_{te}) - Err(\widehat{\theta}^{0}; \epsilon_{te}) \ge \Phi\left(\frac{r/2}{\varphi_{min}} \right) - \Phi\left(\frac{r/2 - \min\{\epsilon_{tr}, \tilde{\epsilon}\}}{\varphi_{min}} \right). \tag{6}$$



(a) Robust error vs ϵ_{tr} (b) Robust error decomposition (c) Intuition in 2D

Figure 4: (a) We set $d = 1000$ and $r = 12$. The robust error as a function of the adversarial training budget ϵ_{tr} for different d/n. (b) The robust error decomposition into susceptibility and standard error as a function of the adversarial budget ϵ_{tr}. Full experimental details can be found in Section C. (c) 2D illustration providing intuition for the linear setting. The effect of adversarial training with directed attacks is captured in the yellow dotted lines: adversarially perturbed training points move closer to the true boundary which in turn tilts the decision boundary more heavily in the wrong direction.

The proof can be found in Appendix A and primarily relies on estimation of singular values of high-dimensional matrices. Note that the theorem holds for any $0 \leq \epsilon_{\text{te}} < \frac{r}{2}$ and hence also directly applies to the standard error by setting $\epsilon_{\text{te}} = 0$. In Figure 3, we empirically confirm the statements of Theorem 3.1 by performing multiple experiments on synthetic datasets as described in Subsection 3.1 with different choices of d/n and ϵ_{tr}. In the first statement, we prove that for small sample-size ($n < d - 1$) noiseless data, almost surely, the robust error increases monotonically with adversarial training budget $\epsilon_{\text{tr}} > 0$. In Figure 3a, we plot the robust error gap between standard and adversarial logistic regression as a function of the adversarial training budget ϵ_{tr} for 5 runs.

The second statement establishes a simplified lower bound on the robust error increase for adversarial training (for a fixed $\epsilon_{\text{tr}} = \epsilon_{\text{te}}$) compared to standard training. In Figures 3a and 3c, we show how the lower bound closely predicts the robust error gap in our synthetic experiments. Furthermore, by the dependence of φ_{min} on the overparameterization ratio d/n, the lower bound on the robust error gap is amplified for large d/n. Indeed, Figure 3c shows how the error gap increases with d/n both theoretically and experimentally. However, when d/n increases above a certain threshold, the gap decreases again, as standard training fails to learn the signal and yields a high error (see Figure 3b).

3.3 Proof intuition

The reason that adversarial training hurts robust generalization is based on an extreme robust vs. standard error trade-off. We now provide intuition for the effect of directed attacks and the low sample regime on the ϵ_{tr}-robust max-ℓ_2-margin solution by decomposing the robust error $\text{Err}(\theta; \epsilon_{\text{te}})$. Notice that ϵ_{te}-robust error $\text{Err}(\theta; \epsilon_{\text{te}})$ can be written as the probability of the union of two events: the event that the classifier based on θ is wrong and the event that the classifier is susceptible to attacks:

$$\text{Err}(\theta; \epsilon_{\text{te}}) = \mathbb{E}_{x,y\sim\mathbb{P}}\left[\mathbb{I}\{yf_\theta(x) < 0\} \vee \max_{x' \in T(x;\epsilon_{\text{te}})} \mathbb{I}\{f_\theta(x)f_\theta(x') < 0\}\right] \leq \text{Err}(\theta; 0) + \text{Susc}(\theta; \epsilon_{\text{te}}) \tag{7}$$

where $\text{Susc}(\theta; \epsilon_{\text{te}})$ is the expectation of the maximization term in Equation 7. $\text{Susc}(\theta; \epsilon_{\text{te}})$ represents the ϵ_{te}-*attack-susceptibility* of a classifier induced by θ and $\text{Err}(\theta; 0)$ its standard error. In our linear setting, we can lower bound Equation 7 by $\text{Err}(\theta; 0) + \frac{1}{2}\text{Susc}(\theta; \epsilon_{\text{te}})$. Hence, Equation 7 suggests that the robust error can only be small if both the standard error and susceptibility are small. In Figure 4b, we plot the decomposition of the robust error in standard error and susceptibility for adversarial logistic regression with increasing ϵ_{tr}. We observe that increasing ϵ_{tr} increases the standard error too drastically compared to the decrease in susceptibility, leading to a drop in robust accuracy. For completeness, in Appendix B, we provide upper and lower bounds for the susceptibility score. We now explain why, in the small-sample size regime, adversarial training with directed attacks 3 may increase standard error to the extent that it dominates the decrease in susceptibility.

A key observation is that the robust max-ℓ_2-margin solution of a dataset $D = \{(x_i, y_i)\}_{i=1}^n$ maximizes the minimum margin that reads $\min_{i\in[n]} y_i\theta^\top(x_i - y_i\epsilon_{\text{tr}}|\theta_{[1]}|e_1)$, where $\theta_{[i]}$ refers to the i-th entry of vector θ. Therefore, it simply corresponds to the max-ℓ_2-margin solution of the dataset shifted towards the decision boundary $D_{\epsilon_{\text{tr}}} = \{(x_i - y_i\epsilon_{\text{tr}}|\widehat{\theta}^{\epsilon_{\text{tr}}}_{[1]}|e_1, y_i)\}_{i=1}^n$. Using this fact, we obtain a closed-form

expression of the (normalized) max-margin solution 4 as a function of ϵ_{tr} that reads

$$\widehat{\theta}^{\epsilon_{\text{tr}}} = \frac{1}{(r - 2\epsilon_{\text{tr}})^2 + 4\tilde{\gamma}^2} \left[r - 2\epsilon_{\text{tr}}, 2\tilde{\gamma}\tilde{\theta} \right], \tag{8}$$

where $\|\tilde{\theta}\|_2 = 1$ and $\tilde{\gamma} > 0$ is a random quantity associated with the max-ℓ_2-margin solution of the $d-1$ dimensional Gaussian inputs orthogonal to the signal direction (see Lemma A.1 in Section A).

In high dimensions, with high probability any two Gaussian random vectors are far apart – in our distributional setting, this corresponds to the vectors being far apart in the non-signal directions. In Figure 4c, we illustrate the phenomenon using a 2D cartoon, where the few samples in the dataset are all far apart in the non-signal direction. We see how shifting the dataset closer to the true decision boundary, may result in a max-margin solution (yellow) that aligns much worse with the ground truth (gray), compared to the estimator learned from the original points (blue). Even though the new (robust max-margin) classifier (yellow) is less susceptible to attacks in the signal dimension, it also uses the signal dimension less. Mathematically, this is reflected in the expression of the max-margin solution in Equation 8: We see that the first (signal) dimension is used less as ϵ_{tr} increases.

3.4 Generality of the results

In this section we discuss how Theorem 3.1 might generalize to other perturbation sets and models.

Signal direction is known The type of additive perturbations used in Theorem 3.1, defined in Equation 3, is explicitly constrained to the direction of the true signal. This choice is reminiscent of corruptions where every possible perturbation in the set is directly targeted at the object to be recognized, such as motion blur of moving objects. Such corruptions are also studied in the context of domain generalization and adaptation (Schneider et al., 2020).

Directed attacks in general, however, may also consist of perturbation sets that are only strongly biased towards the true signal direction. They may find the true signal direction only when the inner maximization is exact. The following corollary extends Theorem 3.1 to small ℓ_1-perturbations

$$T(x; \epsilon) = \{x' = x + \delta \mid \|\delta\|_1 \leq \epsilon\}, \tag{9}$$

for $0 < \epsilon < \frac{r}{2}$ that reflect such attacks. We state the corollary here and give the proof in Appendix A.

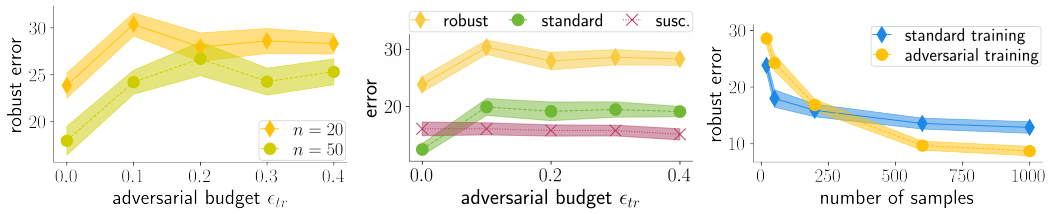
Corollary 3.2. *Theorem 3.1 also holds for 4 with perturbation sets defined in 9.*

The proof uses the fact that the inner maximization effectively results in a sparse perturbation equivalent to the attack resulting from the perturbation set defined in Equation 3.

Other models Motivated by the implicit bias results of (stochastic) gradient descent on the logistic loss, Theorem 3.1 is proven for the max-ℓ_2-margin solution. We would like to conjecture that for the data distribution in Section 3, adversarial training can hurt robust generalization also for other models with zero training error (*interpolators* in short). For example, Adaboost is a widely used algorithm that converges to the max-ℓ_1-margin classifier (Telgarsky, 2013). One might argue that for a sparse ground truth, the max-ℓ_1-margin classifier should (at least in the noiseless case) have the right inductive bias to alleviate large bias in high dimensions. Hence, in many cases the (sparse) max-ℓ_1-margin solution might align with the ground truth for a given dataset. However, we conjecture that even in this case, the *robust* max-ℓ_1-margin solution would be misled to choose a wrong sparse solution. This can be seen with the help of the cartoon illustration in Figure 4c.

4 Real-world experiments

In this section, we demonstrate that the proof intuition of the linear case may generalize to more complex models. Specifically, the insights from Section 3 helped us to identify realistic directed attacks on standard image datasets for which adversarial training hurts robust accuracy in the low sample regime. In what follows, we present experimental results for corruption attacks on the Waterbirds dataset. Due to space constraints, results on the mask attacks on CIFAR-10 can be found in Appendix E. The corresponding experimental details and more results on other additional image datasets (such as the hand gestures dataset) can be found in Appendices D, E and F.



(a) Robust error with increasing ϵ_{tr} (b) Robust error decomposition (c) Robust error vs. #samples

Figure 5: Experiments on the Waterbirds dataset considering the adversarial illumination attack with $\epsilon_{\text{te}} = 0.3$. We plot the mean and standard deviation of the mean of several independent experiments. (a) The robust error increases with larger ϵ_{tr} in the low sample size regime. (b) We set $n = 20$ and plot the robust error decomposition as in Equation 7 with increasing ϵ_{tr}. While the susceptibility decreases slightly, the increase in standard error is much more severe, resulting in an increase in robust error. (c) The robust error of standard training and adversarial training as a function of the number of samples, where the smallest sample size still yields small ($< 10\%$) standard test error for standard training. While adversarial training hurts for small sample sizes, it helps for larger sample sizes. For more experimental details see App. D.

4.1 Datasets and models

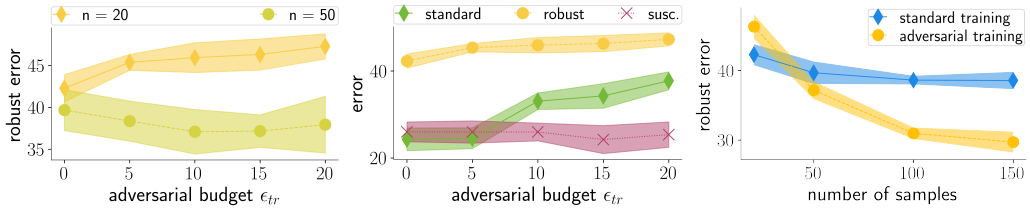
We consider three datasets: the Waterbirds dataset, CIFAR-10 and a hand gesture datasets. Due to space constraints, we describe CIFAR-10 and the hand gesture dataset in Appendix E and F. Apart from CIFAR-10 and the hand gesture dataset, we build a new version of the Waterbirds dataset, consisting of images of water- and landbirds of size 256×256 and labels that distinguish the two types of birds. Using code provided by Sagawa et al. (2020), we construct the dataset as follows: First, we sample equally many water- and landbirds from the CUB-200 dataset (Welinder et al., 2010). Then, we segment the birds and paste them onto a background image that is randomly sampled (without replacement) from the Places-256 dataset (Zhou et al., 2017). Also, following the choice of Sagawa et al. (2020), we use as models a ResNet50 and a ResNet18 that were both pretrained on ImageNet and achieve near perfect standard accuracy. In Appendix D, we complement the results of this section by reporting the results of similar experiments with different architectures.

4.2 Implementation of the directed attacks

In this section, we consider two attacks on the Waterbirds dataset: motion blur and adversarial illumination as depicted in Figure 2. In Appendix E, we also discuss the mask attack, which should mimic occlusions of objects in images that are physically realizable (Eykholt et al., 2018; Wu et al., 2020). On the other hand, motion blur may arise naturally when photographing fast moving objects with a slow shutter speed. Lastly, adversarial illumination may result from adversarial lighting conditions. Next, we describe the motion blur and adversarial illumination attacks in more detail.

Motion blur For the Waterbirds dataset we can implement motion blur attacks on the object (the bird) specifically, a natural corruption that could occur if birds move at speeds that are faster than the shutter speed. The aim is to be robust against all motion blur severity levels up to $M_{max} = 15$. To simulate motion blur, we apply a motion blur filter with a kernel of size M on the segmented bird before we paste it onto the background image. We can change the severity level of the motion blur by increasing the kernel size of the filter. See Appendix D for concrete expressions of the motion blur kernel. Intuitively the worst attack should be the most severe blur, rendering a search over a range of severity superfluous. However, similar to rotations, this is not necessarily true in practice since the training loss on neural networks is generally nonconvex. Therefore, for an exact evaluation of the robust error at test time, we perform a full grid search over all kernel sizes in $[1, 2, \ldots, M_{max}]$. We refer to Figure 2d and Section D for an illustration of our motion blur attack. During training time, we perform an approximate search over kernels with sizes $2i$ for $i = 1, \ldots, M_{max}/2$.

Adversarial illumination As a second attack on the Waterbirds dataset, we consider adversarial illumination. The adversary can darken or brighten the bird without corrupting the background of the image. The attack aims to model images where the object at interest is hidden in shadows or placed against bright light. To compute the adversarial illumination attack, we modify the brightness of the segmented bird by adding a constant $a \in [-\epsilon_{\text{te}}, \epsilon_{\text{te}}]$ to all pixel values, before pasting the bird onto the background image. With an analogous argument as for the adversarial search for motion blur, the



(a) Robust error with increasing ϵ_{tr} (b) Robust error decomposition (c) Robust error vs. #samples

Figure 6: Experiments on the (subsampled) Waterbirds dataset using the motion blur attack. (a) Even though adversarial training hurts robust generalization for low sample size ($n = 20$), it helps for $n = 50$. (b) For $n = 20$, the decomposition of the robust error in standard error and susceptibility as a function of adversarial budget ϵ_{tr}. The increase in standard error is more severe than the drop in susceptibility, leading to a slight increase in robust error. (c) The robust error of standard and adversarial training on settings where the test error after standard training is small as a function of the number of samples. While adversarial training hurts for small sample sizes, it helps for larger sample sizes. For each experiment we plot the mean and standard deviation of the mean of independent experiments. For more experimental details see App. D.

exact evaluation requires an actual search over the interval $[-\epsilon_{\text{te}}, \epsilon_{\text{te}}]$. We find the most adversarial lighting level, i.e. the value of a, by equidistantly partitioning the interval $[-\epsilon_{\text{te}}, \epsilon_{\text{te}}]$ in K steps and performing a full list-search over all steps. See Figure 2c and Appendix D for an illustration of the adversarial illumination attack. We choose $K = 65, 33$ during test and training time respectively.

Adversarial training For all datasets and attacks, we run SGD until convergence on the *robust* cross-entropy loss 2. In each iteration, we search for an adversarial example as described above and update the weights using a gradient with respect to the resulting perturbed example (Goodfellow et al., 2015; Madry et al., 2018). For every experiment, we choose the learning rate and weight decay parameters that minimize the robust error on a hold-out dataset.

4.3 Adversarial training can hurt robust generalization

We now present our experimental results on the Waterbirds dataset for both motion blur and adversarial illumination attacks. First of all, Figure 5a and 6a show that the phenomenon characterized in the linear setting by Theorem 3.1 also occurs for directed attacks on the Waterbirds dataset: as we increase the adversarial training budget ϵ_{tr} starting from zero (standard training), the robust error monotonically increases.

Furthermore, to gain intuition as described in Section 3.3, we also plot the robust error decomposition (Equation 7) consisting of the standard error and susceptibility in Figure 5b and 6b. Recall that we measure susceptibility as the fraction of data points in the test set for which the classifier predicts a different class under an adversarial attack. As in our linear example, we observe an increase in robust error despite a slight drop in susceptibility, because of the more severe increase in standard error. Moreover, Figures 1 and 6c show that analogous to our linear example, this phenomenon is specific to the low sample regime: for large sample size adversarial training outperforms standard training as expected. Note again that even the smallest sample size is large enough to yield a standard test error $< 10\%$ for standard training. Similar experiments for CIFAR-10 can be found in Appendix E. Finally, we empirically confirm in Appendix D.8 that our phenomenon is specific to directed attacks: for undirected attacks such as bounded $\ell_\infty-$and ℓ_2-ball perturbations, adversarial training helps robust generalization also in the low sample size regime.

4.4 Discussion

We now discuss how different algorithmic choices, motivated by related work, might affect how adversarial training hurts robust generalization.

Catastrophic overfitting Often the worst-case perturbation during adversarial training is found using an approximate algorithm such as SGD. It is common belief that using the strongest attack (in the motion blur case, full grid search) during training also results in better robust generalization. In particular, the literature on catastrophic overfitting shows that weaker attacks during training lead to bad performance on stronger attacks during testing (Wong et al., 2020; Andriushchenko & Flammarion, 2020; Li et al., 2021). Our results suggests the opposite in the low sample size regime for directed attacks: the weaker the attack during training, the better adversarial training performs.

Robust overfitting Recent work observes empirically (Rice et al., 2020) and theoretically (Sanyal et al., 2020; Donhauser et al., 2021), that perfectly minimizing the adversarial loss during training might in fact be suboptimal for robust generalization; that is, classical regularization techniques might lead to higher robust accuracy. This phenomenon is often referred to as robust overfitting. May the phenomenon be mitigated using standard regularization techniques? In Appendix D we shed light on this question and show that adversarial training hurts robust generalization even when standard regularization methods such as early stopping are used.

5 Related work

Robust and non-robust useful features In the words of Ilyas et al. (2019) and Springer et al. (2021) we can describe the intuition behind "our phenomenon" as follows: for directed attacks, all robust features become less useful, but adversarial training uses robust features more. In the small sample-size regime, $n < d - 1$ in particular, robust learning assigns too much weight on the robust (possibly non-useful) features that then dominate the non-robust (but useful)features. Even though they define these concepts, they don't make our statement, but show that adversarial training reduces the reliance on non-robust but possibly useful features.

Small sample size and robustness A direct consequence of Theorem 3.1 is that in order to achieve the same robust error as standard training, adversarial training requires more samples. This statement might remind the reader of sample complexity results for robust generalization in Schmidt et al. (2018); Yin et al. (2019); Khim & Loh (2018). While those results compare sample complexity bounds for standard vs. robust error, our theorem statement compares two algorithms, standard vs. adversarial training, with respect to the robust error.

Trade-off between standard and robust error Many papers observed that even though adversarial training decreases robust error compared to standard training, it may lead to an increase in standard test error Madry et al. (2018); Zhang et al. (2019). For example, Tsipras et al. (2019); Zhang et al. (2019); Javanmard et al. (2020); Dobriban et al. (2020); Chen et al. (2020) study settings where the Bayes optimal robust classifier is not equal to the Bayes optimal (standard) classifier (i.e. the perturbations are inconsistent or the dataset is non-separable). Raghunathan et al. (2020) study consistent perturbations, as in our paper, and prove that for small sample size, fitting adversarial examples can increase standard error even in the absence of noise. Empirically, Dong et al. (2021); Mendonça et al. (2022) show that for ℓ_p-attacks low-quality data might be the main cause of the trade-off. While aforementioned works focus on the decrease in standard error, we prove that for directed attacks, in the small sample regime adversarial training may in fact increase *robust error*.

Mitigation of the trade-off A long line of work has proposed procedures to mitigate the trade-off between robust and standard accuracy. For example Alayrac et al. (2019); Carmon et al. (2019); Zhai et al. (2019); Raghunathan et al. (2020) study robust self training, which leverages a large set of unlabelled data, while Lee et al. (2020); Lamb et al. (2019); Xu et al. (2020) use data augmentation by interpolation. Ding et al. (2020); Balaji et al. (2019); Cheng et al. (2020) on the other hand propose to use adaptive perturbation budgets ϵ_{tr} that vary across inputs. The intuition behind our theoretical analysis suggests that the standard mitigation procedures for imperceptible perturbations may not work for perceptible directed attacks, because all relevant features are non-robust. We leave a thorough empirical study as interesting future work.

6 Summary and future work

This paper aims to caution the practitioner against blindly following current widespread practices to increase the robust performance of machine learning models. Specifically, adversarial training is currently recognized to be one of the most effective defense mechanisms for ℓ_p-perturbations, significantly outperforming robust performance of standard training. However, we prove that in the low sample size regime this common wisdom is not applicable for consistent directed attacks, which efficiently focus their attack budget to target the ground truth class information. In terms of follow-up work on directed attacks in the low sample regime, there are some concrete questions that would be interesting to explore. For example, as discussed in Section 5, it would be useful to test whether some methods to mitigate the standard accuracy vs. robustness trade-off would also relieve the perils of adversarial training for directed attacks. Further, we hypothesize that when few samples are available, one should avoid training with attacks that may heavily reduce class information, independently of

the attacks at test time. If this hypothesis were confirmed, it would break with yet another general rule that the best defense perturbation type should always match the attack during evaluation.

REFERENCES

Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? *Advances in Neural Information Processing Systems*, pp. 12214–12223, 2019.

Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 2020.

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In Zhi-Hua Zhou (ed.), *The 30th International Joint Conference on Artificial Intelligence*, pp. 4312–4321. International Joint Conferences on Artificial Intelligence Organization, 2021.

Yogesh Balaji, Tom Goldstein, and Judy Hoffman. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. *arXiv preprint arXiv:1910.08051*, 2019.

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. In *The 33rd International Conference on Neural Information Processing Systems*, pp. 11192–11203, 2019.

Lin Chen, Yifei Min, Mingrui Zhang, and Amin Karbasi. More data can expand the generalization gap between adversarially robust and standard models. In *The 36th International Conference on Machine Learning*, pp. 1670–1680, 2020.

Minhao Cheng, Qi Lei, Pin-Yu Chen, Inderjit Dhillon, and Cho-Jui Hsieh. Cat: Customized adversarial training for improved robustness. *arXiv preprint arXiv:2002.06789*, 2020.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *The 7th International Conference on Learning Theory*, pp. 1305–1338, 2020.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *The 37th International Conference on Machine Learning*, pp. 2206–2216, 2020.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. In *The 8th International Conference on Learning Representations*, 2020.

Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

Chengyu Dong, Liyuan Liu, and Jingbo Shang. Data quality matters for adversarial training: An empirical study, 2021.

Konstantin Donhauser, Alexandru Tifrea, Michael Aerni, Reinhard Heckel, and Fanny Yang. Interpolation can hurt robust generalization even when there is no noise. *The 36th conference on Advances in Neural Information Processing Systems*, 2021.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *The 36th International Conference on Machine Learning*, pp. 1802–1811, 2019.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1625–1634, 2018.

Amin Ghiasi, Ali Shafahi, and Tom Goldstein. Breaking certified defenses: semantic adversarial examples with spoofed robustness certificates. In *The 6th International Conference on Learning Representations*, 2019.

Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *The 3th International Conference on Learning Representations*, pp. 1–10, 2015.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *The 33rd conference on Advances in Neural Information Processing Systems*, pp. 125–136, 2019.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pp. 2034–2078, 2020.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *The 32nd Conference on Learning Theory*, pp. 1772–1798, 2019.

Daniel Kang, Yi Sun, Tom Brown, Dan Hendrycks, and Jacob Steinhardt. Transfer of adversarial robustness between perturbation types. *arXiv e-prints*, pp. arXiv–1905, 2019.

Justin Khim and Po-Ling Loh. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *The 9th International Conference on Learning Representation*, 2021.

Alex Lamb, Vikas Verma, Juho Kannala, and Yoshua Bengio. Interpolated adversarial training: Achieving robust neural networks without sacrificing too much accuracy. In *The 12th ACM Workshop on Artificial Intelligence and Security*, pp. 95–103, 2019.

Saehyung Lee, Hyungyu Lee, and Sungroh Yoon. Adversarial vertex mixup: Toward better adversarially robust generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Bai Li, Shiqi Wang, Suman Jana, and Lawrence Carin. Towards understanding fast adversarial training. *arXiv preprint arXiv:2006.03089*, 2021.

Yan Li, Ethan X.Fang, Huan Xu, and Tuo Zhao. Implicit bias of gradient descent based adversarial training on separable data. In *The 8th International Conference on Learning Representations*, 2020.

Wei-An Lin, Chun Pong Lau, Alexander Levine, Rama Chellappa, and Soheil Feizi. Dual manifold adversarial robustness: Defense against lp and non-lp adversarial attacks. In *The 34th conference on Advances in Neural Information Processing Systems*, pp. 3487–3498, 2020.

Chen Liu, Mathieu Salzmann, Tao Lin, Ryota Tomioka, and Sabine Süsstrunk. On the loss landscape of adversarial training: Identifying challenges and how to overcome them. In *The 35th conference on Advances in Neural Information Processing Systems*, pp. 21476–21487, 2020.

Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. Towards imperceptible and robust adversarial example attacks against neural networks. In *The 32nd AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *The 6th International Conference on Learning Representations*, 2018.

Tomás Mantecón, Carlos R. del Blanco, Fernando Jaureguizar, and Narciso García. A real-time gesture recognition system using near-infrared imagery. *PLOS ONE*, pp. 1–17, 2019.

Marcele OK Mendonça, Javier Maroto, Pascal Frossard, and Paulo SR Diniz. Adversarial training with informed data selection. In *The 30th European Signal Processing Conference (EUSIPCO)*, pp. 608–612, 2022.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *The IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2574–2582, 2016.

Abdullah Mujahid, Mazhar Javed Awan, Awais Yasin, Mazin Abed Mohammed, Robertas Damaševičius, Rytis Maskeliūnas, and Karrar Hameed Abdulkareem. Real-time hand gesture recognition based on deep learning yolov3 model. *Applied Sciences*, 2021.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3051–3059, 2019.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *The 33d conference on Advances in Neural Information Processing Systems*, pp. 11611–11622, 2019.

Munir Oudah, Ali Al-Naji, and Javaan Chahl. Hand gesture recognition based on computer vision: A review of techniques. *Journal of Imaging*, 2020.

Huy Phan. huyvnphan/pytorch_cifar10, 1 2021.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *The 37th International Conference on Machine Learning*, pp. 7909–7919, 2020.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *The 37th International Conference on Machine Learning*, pp. 8093–8104, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *The 7th International Conference on Learning Representations*, 2020.

Amartya Sanyal, Puneet K Dokania, Varun Kanade, and Philip Torr. How benign is benign overfitting? In *The 8th International Conference on Learning Representations*, 2020.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *The 32nd conference Advances in Neural Information Processing Systems*, 2018.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *The 34th conference on Advances in Neural Information Processing Systems*, pp. 11539–11551, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, pp. 1–57, 2018.

Jacob M Springer, Melanie Mitchell, and Garrett T Kenyon. Adversarial perturbations are not so weird: Entanglement of robust and non-robust features in neural network classifiers. *arXiv preprint arXiv:2102.05110*, 2021.

David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6967–6987, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *The 2nd International Conference on Learning Representations*, 2014.

Matus Telgarsky. Margins, shrinkage, and boosting. In *The 30th International Conference on Machine Learning*, pp. 307–315, 2013.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *The 7th International Conference on Learning Representations*, 2019.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical report, California Institute of Technology, 2010.

Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *The 8th International Conference on Learning Representations*, 2020.

Tong Wu, Liang Tong, and Yevgeniy Vorobeychik. Defending against physically realizable attacks on image classification. In *The 8th International Conference on Learning Representations*, 2020.

Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *The AAAI Conference on Artificial Intelligence*, pp. 6502–6509, 2020.

Shuai Yang, Prashan Premaratne, and Peter Vial. Hand gesture recognition: An overview. In *The 5th IEEE International Conference on Broadband Network Multimedia Technology*, pp. 63–69, 2013.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *The 36th International conference on machine learning*, pp. 7085–7094, 2019.

Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *The 36th International Conference on Machine Learning*, pp. 7472–7482, 2019.

Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1039–1048, 2020.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Jianli Zhou, Chao Liang, and Jun Chen. Manifold projection for adversarial defense on face recognition. In *The 16th European Conference on Computer Vision*, pp. 288–305, 2020.