Adversarially Robust Generalization Requires More Data

Ludwig Schmidt
UC Berkeley
ludwig@berkeley.edu

Shibani Santurkar
MIT
shibani@mit.edu

Dimitris Tsipras
MIT
tsipras@mit.edu

Kunal Talwar
Google Brain
kunal@google.com

Aleksander Mądry
MIT
madry@mit.edu

Abstract

Machine learning models are often susceptible to adversarial perturbations of their inputs. Even small perturbations can cause state-of-the-art classifiers with high "standard" accuracy to produce an incorrect prediction with high confidence. To better understand this phenomenon, we study adversarially robust learning from the viewpoint of generalization. We show that already in a simple natural data model, the sample complexity of robust learning can be significantly larger than that of "standard" learning. This gap is information theoretic and holds irrespective of the training algorithm or the model family. We complement our theoretical results with experiments on popular image classification datasets and show that a similar gap exists here as well. We postulate that the difficulty of training robust classifiers stems, at least partially, from this inherently larger sample complexity.

1 Introduction

Modern machine learning models achieve high accuracy on a broad range of datasets, yet can easily be misled by small perturbations of their input. While such perturbations are often simple noise to a human or even imperceptible, they cause state-of-the-art models to misclassify their input with high confidence. This phenomenon has first been studied in the context of secure machine learning for spam filters and malware classification [7, 16, 35]. More recently, researchers have demonstrated the phenomenon under the name of *adversarial examples* in image classification [21, 51], question answering [28], voice recognition [12, 13, 49, 62], and other domains (for instance, see [2, 4, 14, 22, 25, 26, 32, 60]). Overall, the existence of such adversarial examples raises concerns about the robustness of current classifiers. As we increasingly deploy machine learning systems in safety- and security-critical environments, it is crucial to understand the robustness properties of our models in more detail.

A growing body of work is exploring this robustness question from the security perspective by proposing *attacks* (methods for crafting adversarial examples) and *defenses* (methods for making classifiers robust to such perturbations). Often, the focus is on deep neural networks, e.g., see [11, 24, 36, 37, 41, 47, 53, 59]. While there has been success with robust classifiers on simple datasets [31, 36, 44, 48], more complicated datasets still exhibit a large gap between "standard" and robust accuracy [3, 11]. An implicit assumption underlying most of this work is that the same training dataset that enables good standard accuracy also suffices to train a robust model. However, it is unclear if this assumption is valid.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

So far, the *generalization* aspects of adversarially robust classification have not been thoroughly investigated. Since adversarial robustness is a learning problem, the statistical perspective is of integral importance. A key observation is that adversarial examples are not at odds with the standard notion of generalization as long as they occupy only a small total measure under the data distribution. So to achieve adversarial robustness, a classifier must generalize in a stronger sense. We currently do not have a good understanding of how such a stronger notion of generalization compares to standard "benign" generalization, i.e., without an adversary.

In this work, we address this gap and explore the statistical foundations of adversarially robust generalization. We focus on sample complexity as a natural starting point since it underlies the core question of when it is possible to learn an adversarially robust classifier. Concretely, we pose the following question:

> *How does the sample complexity of standard generalization compare to that of adversarially robust generalization?*

Put differently, we ask if a dataset that allows for learning a good classifier also suffices for learning a robust one. To study this question, we analyze robust generalization in two distributional models. By focusing on specific distributions, we can establish information-theoretic lower bounds and describe the exact sample complexity requirements for generalization. We find that even for a simple data distribution such as a mixture of two class-conditional Gaussians, the sample complexity of robust generalization is significantly larger than that of standard generalization. Our lower bound holds for *any* model and learning algorithm. Hence no amount of algorithmic ingenuity is able to overcome this limitation.

In spite of this negative result, simple datasets such as MNIST have recently seen significant progress in terms of adversarial robustness [31, 36, 44, 48]. The most robust models achieve accuracy around 90% against large ℓ_∞-perturbations. To better understand this discrepancy with our first theoretical result, we also study a second distributional model with binary features. This binary data model has the same standard generalization behavior as the previous Gaussian model. Moreover, it also suffers from a significantly increased sample complexity whenever one employs *linear* classifiers to achieve adversarially robust generalization. Nevertheless, a slightly non-linear classifier that utilizes thresholding turns out to recover the smaller sample complexity of standard generalization. Since MNIST is a mostly binary dataset, our result provides evidence that ℓ_∞-robustness on MNIST is significantly easier than on other datasets. Moreover, our results show that distributions with similar sample complexity for standard generalization can still exhibit considerably different sample complexity for robust generalization.

To complement our theoretical results, we conduct a range of experiments on MNIST, CIFAR10, and SVHN. By subsampling the datasets at various rates, we study the impact of sample size on adversarial robustness. When plotted as a function of training set size, our results show that the standard accuracy on SVHN indeed plateaus well before the adversarial accuracy reaches its maximum. On MNIST, explicitly adding thresholding to the model during training significantly reduces the sample complexity, similar to our upper bound in the binary data model. On CIFAR10, the situation is more nuanced because there are no known approaches that achieve more than 50% accuracy even against a mild adversary. But as we show below, there is clear evidence for overfitting in the current state-of-the-art methods.

Overall, our results suggest that current approaches may be unable to attain higher adversarial accuracy on datasets such as CIFAR10 for a fundamental reason: the dataset may not be large enough to train a standard convolutional network robustly. Moreover, our lower bounds illustrate that the existence of adversarial examples should not necessarily be seen as a shortcoming of specific classification methods. Already in a simple data model, adversarial examples *provably* occur for any learning approach, even when the classifier already achieves high standard accuracy. So while vulnerability to adversarial ℓ_∞-perturbations might seem counter-intuitive at first, in some regimes it is an unavoidable consequence of working in a statistical setting.

1.1 A motivating example: Overfitting on CIFAR10

Before we describe our main results, we briefly highlight the importance of generalization for adversarial robustness via two experiments on MNIST and CIFAR10. In both cases, our goal is to learn a classifier that achieves good test accuracy even under ℓ_∞-bounded perturbations. We follow

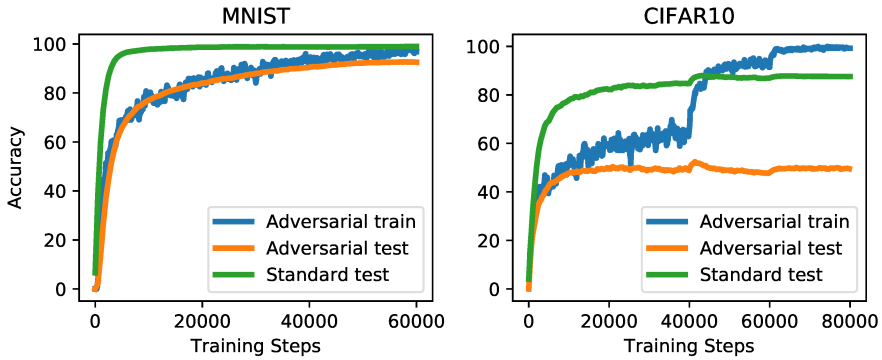

Figure 1: Classification accuracies for robust optimization on MNIST and CIFAR10. In both cases, we trained standard convolutional networks to be robust to ℓ_∞-perturbations of the input. On MNIST, the robust test error closely tracks the corresponding training error and the model achieves high robust accuracy. On CIFAR10, the model still achieves a good natural (non-adversarial) test error, but there is a significant generalization gap for the robust accuracy. This phenomenon motivates our study of adversarially robust generalization.

the standard robust optimization approach [6, 36, 54] – also known as adversarial training [21, 51] – and (approximately) solve the saddle point problem

$$\min_{\theta} \mathbb{E}_{x} \left[\max_{\|x'-x\|_\infty \leq \varepsilon} \text{loss}(\theta, x') \right]$$

via stochastic gradient descent over the model parameters θ. We utilize projected gradient descent for the inner maximization problem over allowed perturbations of magnitude ε (see [36] for details). Figure 1 displays the training curves for three quantities: (i) adversarial training error, (ii) adversarial test error, and (iii) standard test error.

The results show that on MNIST, robust optimization is able to learn a model with around 90% adversarial accuracy and a relatively small gap between training and test error. However, CIFAR10 offers a different picture. Here, the model (a wide residual network [61]) is still able to fully fit the training set even against an adversary, but the generalization gap is significantly larger. The model only achieves 47% adversarial test accuracy, which is about 50% lower than its training accuracy.[1] Moreover, the standard test accuracy is about 87%, so the failure of generalization indeed primarily occurs in the context of adversarial robustness. This failure may be surprising particularly since properly tuned convolutional networks rarely overfit much on standard vision datasets.

1.2 Outline of the paper

In the next section, we describe our main theoretical results at a high level. Section 3 complements these results with experiments. We discuss related works in Section 4 and conclude in Section 5. Due to space constraints, a longer discussion of related work, several open questions, and all proofs are deferred to the appendix in the supplementary material.

2 Theoretical Results

Our theoretical results concern statistical aspects of adversarially robust classification. In order to understand how properties of data affect the number of samples needed for robust generalization, we study two concrete distributional models.

While our two data models are clearly much simpler than the image datasets currently being used in the experimental work on ℓ_∞-robustness, we believe that the simplicity of our models is a strength in this context. The fact that we can establish a separation between standard and robust generalization already in our Gaussian data model is evidence that the existence of adversarial examples for neural

[1] We remark that this accuracy is still currently the best published robust accuracy on CIFAR10 [3]. For instance, contemporary approaches to architecture tuning do not yield better robust accuracies [15].

networks should not come as a surprise. The same phenomenon (i.e., classifiers with just enough samples for high standard accuracy *necessarily* being vulnerable to ℓ_∞- attacks) already occurs in much simpler settings such as a mixture of two Gaussians. Note that more complicated distributional setups that can "simulate" the Gaussian model directly inherit our lower bounds.

In addition, conclusions from our simple models also transfer to real datasets. As we describe in the subsection on the Bernoulli model, the benefits of the thresholding layer predicted by our theoretical analysis do indeed appear in experiments on MNIST as well. Since multiple defenses against adversarial examples have been primarily evaluated on MNIST [31, 44, 48], it is important to note that ℓ_∞-robustness on MNIST is a particularly easy case: adding a simple thresholding layer directly yields nearly state-of-the-art robustness against moderately strong adversaries ($\varepsilon = 0.1$), without any further changes to the model architecture or training algorithm.

2.1 The Gaussian model

Our first data model is a mixture of two spherical Gaussians with one component per class.

Definition 1 (Gaussian model)**.** *Let $\theta^\star \in \mathbb{R}^d$ be the per-class mean vector and let $\sigma > 0$ be the variance parameter. Then the $(\theta^\star, \sigma)$-Gaussian model is defined by the following distribution over $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$: First, draw a label $y \in \{\pm 1\}$ uniformly at random. Then sample the data point $x \in \mathbb{R}^d$ from $\mathcal{N}(y \cdot \theta^\star, \sigma^2 I)$.*

While not explicitly specified in the definition, we will use the Gaussian model in the regime where the norm of the vector $\theta^\star$ is approximately $\sqrt{d}$. Hence the main free parameter for controlling the difficulty of the classification task is the variance σ^2, which controls the amount of overlap between the two classes.

To contrast the notions of "standard" and "robust" generalization, we briefly recap a standard definition of classification error.

Definition 2 (Classification error)**.** *Let $\mathcal{P} : \mathbb{R}^d \times \{\pm 1\} \to \mathbb{R}$ be a distribution. Then the classification error β of a classifier $f : \mathbb{R}^d \to \{\pm 1\}$ is defined as $\beta = \mathbb{P}_{(x,y)\sim\mathcal{P}}[f(x) \neq y]$.*

Next, we define our main quantity of interest, which is an adversarially robust counterpart of the above classification error. Instead of counting misclassifications under the data distribution, we allow a bounded worst-case perturbation before passing the perturbed sample to the classifier.

Definition 3 (Robust classification error)**.** *Let $\mathcal{P} : \mathbb{R}^d \times \{\pm 1\} \to \mathbb{R}$ be a distribution and let $\mathcal{B} : \mathbb{R}^d \to \mathscr{P}(\mathbb{R}^d)$ be a perturbation set.[2] Then the $\mathcal{B}$-robust classification error β of a classifier $f : \mathbb{R}^d \to \{\pm 1\}$ is defined as $\beta = \mathbb{P}_{(x,y)\sim\mathcal{P}}[\exists\, x' \in \mathcal{B}(x) \,:\, f(x') \neq y]$.*

Since ℓ_∞-perturbations have recently received a significant amount of attention, we focus on robustness to ℓ_∞-bounded adversaries in our work. For this purpose, we define the perturbation set $\mathcal{B}^\varepsilon_\infty(x) = \{x' \in \mathbb{R}^d \mid \|x' - x\|_\infty \leq \varepsilon\}$. To simplify notation, we refer to robustness with respect to this set also as ℓ^ε_∞-robustness. As we remark in the discussion section, understanding generalization for other measures of robustness (ℓ_2, rotations, etc.) is an important direction for future work.

Standard generalization. The Gaussian model has one parameter for controlling the difficulty of learning a good classifier. In order to simplify the following bounds, we study a regime where it is possible to achieve good *standard* classification error from a single sample.[3] As we will see later, this also allows us to calibrate our two data models to have comparable standard sample complexity.

Concretely, we prove the following theorem, which is a direct consequence of Gaussian concentration. Note that in this theorem we use a *linear classifier*: for a vector w, the linear classifier $f_w : \mathbb{R}^d \to \{\pm 1\}$ is defined as $f_w(x) = \mathrm{sgn}(\langle w, x\rangle)$.

Theorem 4. *Let (x, y) be drawn from a $(\theta^\star, \sigma)$-Gaussian model with $\|\theta^\star\|_2 = \sqrt{d}$ and $\sigma \leq c \cdot d^{1/4}$ where c is a universal constant. Let $\widehat{w} \in \mathbb{R}^d$ be the vector $\widehat{w} = y \cdot x$. Then with high probability, the linear classifier $f_{\widehat{w}}$ has classification error at most 1%.*

[2] We write $\mathscr{P}(\mathbb{R}^d)$ to denote the power set of $\mathbb{R}^d$, i.e., the set of subsets of $\mathbb{R}^d$.

[3] We remark that it is also possible to study a more general setting where standard generalization requires a larger number of samples.

To minimize the number of parameters in our bounds, we have set the error probability to 1%. By tuning the model parameters appropriately, it is possible to achieve a vanishingly small error probability from a single sample (see Corollary 19 in Appendix D.1).

Robust generalization. As we just demonstrated, we can easily achieve *standard* generalization from only a single sample in our Gaussian model. We now show that achieving a low ℓ_∞-*robust* classification error requires significantly more samples. To this end, we begin with a natural strengthening of Theorem 4 and prove that the (class-weighted) sample mean can also be a robust classifier (given sufficient data).

Theorem 5. *Let $(x_1, y_1), \ldots, (x_n, y_n)$ be drawn i.i.d. from a $(\theta^\star, \sigma)$-Gaussian model with $\|\theta^\star\|_2 = \sqrt{d}$ and $\sigma \leq c_1 d^{1/4}$. Let $\widehat{w} \in \mathbb{R}^d$ be the weighted mean vector $\widehat{w} = \frac{1}{n}\sum_{i=1}^n y_i x_i$. Then with high probability, the linear classifier $f_{\widehat{w}}$ has ℓ_∞^ε-robust classification error at most 1% if*

$$n \;\geq\; \begin{cases} 1 & \text{for} \;\; \varepsilon \leq \frac{1}{4} d^{-1/4} \\ c_2\, \varepsilon^2 \sqrt{d} & \text{for} \;\; \frac{1}{4} d^{-1/4} \leq \varepsilon \leq \frac{1}{4} \end{cases} .$$

We refer the reader to Corollary 22 in Appendix D.1 for the details. As before, c_1 and c_2 are two universal constants. Overall, the theorem shows that it is possible to learn an ℓ_∞^ε-robust classifier in the Gaussian model as long as ε is bounded by a small constant and we have a large number of samples.

Next, we show that this significantly increased sample complexity is necessary. Our main theorem establishes a lower bound for *all* learning algorithms, which we formalize as functions from data samples to binary classifiers. In particular, the lower bound applies not only to learning linear classifiers.

Theorem 6. *Let g_n be any learning algorithm, i.e., a function from n samples to a binary classifier f_n. Moreover, let $\sigma = c_1 \cdot d^{1/4}$, let $\varepsilon \geq 0$, and let $\theta \in \mathbb{R}^d$ be drawn from $\mathcal{N}(0, I)$. We also draw n samples from the (θ, σ)-Gaussian model. Then the expected ℓ_∞^ε-robust classification error of f_n is at least $(1 - 1/d)\frac{1}{2}$ if*

$$n \;\leq\; c_2 \frac{\varepsilon^2 \sqrt{d}}{\log d} .$$

The proof of the theorem can be found in Corollary 23 (Appendix D.2). It is worth noting that the classification error $1/2$ in the lower bound is tight. A classifier that always outputs a fixed prediction trivially achieves perfect robustness on one of the two classes and hence robust accuracy $1/2$.

Comparing Theorems 5 and 6, we see that the sample complexity n required for robust generalization is bounded as

$$\frac{c\varepsilon^2\sqrt{d}}{\log d} \;\leq\; n \;\leq\; c'\varepsilon^2\sqrt{d} \, .$$

Hence the lower bound is nearly tight in our regime of interest. When the perturbation has constant ℓ_∞-norm, the sample complexity of robust generalization is larger than that of standard generalization by $\sqrt{d}$, i.e., *polynomial* in the problem dimension. This shows that for high-dimensional problems, adversarial robustness can provably require a significantly larger number of samples.

Finally, we remark that our lower bound applies also to a more restricted adversary. Our proof uses only a single adversarial perturbation per class. As a result, the lower bound provides *transferable* adversarial examples and applies to worst-case distribution shifts without a classifier-adaptive adversary. We refer the reader to Section 5 for a more detailed discussion.

2.2 The Bernoulli model

As mentioned in the introduction, simpler datasets such as MNIST have recently seen significant progress in terms of ℓ_∞-robustness. We now investigate a possible mechanism underlying these advances. To this end, we study a second distributional model that highlights how the data distribution can significantly affect the achievable robustness. The second data model is defined on the hypercube $\{\pm 1\}^d$, and the two classes are represented by opposite vertices of that hypercube. When sampling a datapoint for a given class, we flip each bit of the corresponding class vertex with a certain probability. This data model is inspired by the MNIST dataset because MNIST images are close to binary (many pixels are almost fully black or white).

Definition 7 (Bernoulli model)**.** *Let $\theta^\star \in \{\pm 1\}^d$ be the per-class mean vector and let $\tau > 0$ be the class bias parameter. Then the $(\theta^\star, \tau)$-Bernoulli model is defined by the following distribution over $(x, y) \in \{\pm 1\}^d \times \{\pm 1\}$: First, draw a label $y \in \{\pm 1\}$ uniformly at random from its domain. Then sample the data point $x \in \{\pm 1\}^d$ by sampling each coordinate x_i from the distribution*

$$x_i \;=\; \begin{cases} y \cdot \theta_i^\star & \text{with probability } 1/2 + \tau \\ -y \cdot \theta_i^\star & \text{with probability } 1/2 - \tau \end{cases} .$$

As in the previous subsection, the model has one parameter for controlling the difficulty of learning. A small value of τ makes the samples less correlated with their respective class vectors and hence leads to a harder classification problem. Note that both the Gaussian and the Bernoulli model are defined by simple sub-Gaussian distributions. Nevertheless, we will see that they differ significantly in terms of robust sample complexity.

Standard generalization. As in the Gaussian model, we first calibrate the distribution so that we can learn a classifier with good *standard* accuracy from a single sample.[4] The following theorem is a direct consequence of the fact that bounded random variables exhibit sub-Gaussian concentration.

Theorem 8. *Let (x, y) be drawn from a $(\theta^\star, \tau)$-Bernoulli model with $\tau \;\geq\; c \cdot d^{-1/4}$ where c is a universal constant. Let $\widehat{w} \in \mathbb{R}^d$ be the vector $\widehat{w} = y \cdot x$. Then with high probability, the linear classifier $f_{\widehat{w}}$ has classification error at most 1%.*

To simplify the bound, we have set the error probability to be 1% as in the Gaussian model. We refer the reader to Corollary 28 in Appendix F.1 for the proof.

Robust generalization. Next, we investigate the sample complexity of robust generalization in our Bernoulli model. For *linear* classifiers, a small robust classification error again requires a large number of samples:

Theorem 9. *Let g_n be a linear classifier learning algorithm, i.e., a function from n samples to a linear classifier f_n. Suppose that we choose $\theta^\star$ uniformly at random from $\{\pm 1\}^d$ and draw n samples from the $(\theta^\star, \tau)$-Bernoulli model with $\tau = c_1 \cdot d^{-1/4}$. Moreover, let $\varepsilon < 3\tau$ and $0 < \gamma < 1/2$. Then the expected ℓ_∞^ε-robust classification error of f_n is at least $\frac{1}{2} - \gamma$ if*

$$n \;\leq\; c_2 \frac{\varepsilon^2 \gamma^2 d}{\log d/\gamma} .$$

We defer the proof to Appendix F.2. At first, the lower bound for linear classifiers might suggest that ℓ_∞-robustness requires an inherently larger sample complexity here as well. However, in contrast to the Gaussian model, non-linear classifiers can achieve a significantly improved robustness. In particular, consider the following thresholding operation $T : \mathbb{R}^d \to \mathbb{R}^d$ which is defined element-wise as

$$T(x)_i \;=\; \begin{cases} +1 & \text{if } x_i \geq 0 \\ -1 & \text{otherwise} \end{cases} .$$

It is easy to see that for $\varepsilon < 1$, the thresholding operator undoes the action of any ℓ_∞-bounded adversary, i.e., we have $T(\mathcal{B}_\infty^\varepsilon(x)) = \{x\}$ for any $x \in \{\pm 1\}^d$. Hence we can combine the thresholding operator with the classifier learned from a single sample to get the following upper bound.

Theorem 10. *Let (x, y) be drawn from a $(\theta^\star, \tau)$-Bernoulli model with $\tau \;\geq\; c \cdot d^{-1/4}$ where c is a universal constant. Let $\widehat{w} \in \mathbb{R}^d$ be the vector $\widehat{w} = yx$. Then with high probability, the classifier $f_{\widehat{w}} \circ T$ has ℓ_∞^ε-robust classification error at most 1% for any $\varepsilon < 1$.*

This theorem shows a stark contrast to the Gaussian case. Although both models have similar sample complexity for standard generalization, there is a $\sqrt{d}$ gap between the ℓ_∞-robust sample complexity for the Bernoulli and Gaussian models. This discrepancy provides evidence that robust generalization requires a more nuanced understanding of the data distribution than standard generalization.

[4] To be precise, the two distributions have comparable sample complexity for standard generalization in the regime where $\sigma \approx \tau^{-1}$.

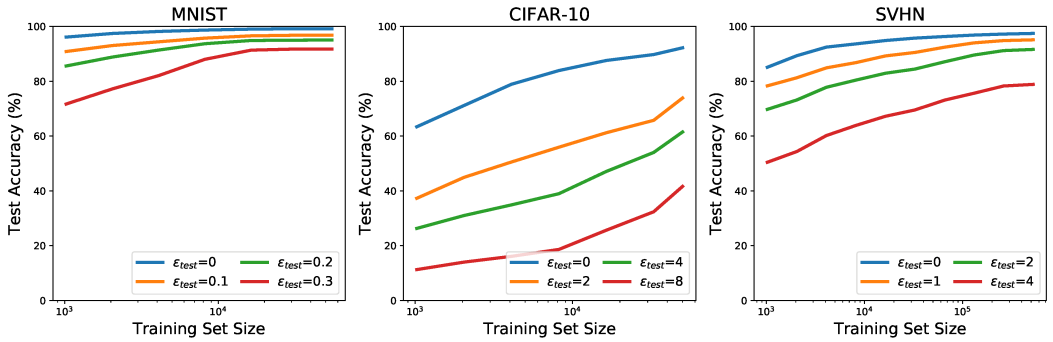

Figure 2: Adversarially robust generalization performance as a function of training data size for ℓ_∞ adversaries on the MNIST, CIFAR-10 and SVHN datasets. For each choice of training set size and ε_{test}, we plot the best performance achieved over ε_{train} and network capacity. This clearly shows that achieving a certain level of adversarially robust generalization requires significantly more samples than achieving the same level of standard generalization.

In isolation, the thresholding step might seem specific to the Bernoulli model studied here. However, our experiments in Section 3 show that an explicit thresholding layer also significantly improves the sample complexity of training a robust neural network on MNIST. We conjecture that the effectiveness of thresholding is behind many of the successful defenses against adversarial examples on MNIST (for instance, see Appendix C in [36]).

3 Experiments

We complement our theoretical results by performing experiments on multiple common datasets. We consider standard convolutional neural networks and train models on datasets of varying complexity. Specifically, we study the MNIST [34], CIFAR-10 [33], and SVHN [40] datasets. We use a simple convolutional architecture for MNIST, a standard ResNet model [23] for CIFAR-10, and a wider ResNet [61] for SVHN. We perform robust optimization to train our classifiers on perturbations generated by projected gradient descent. Appendix G provides additional details for our experiments.

Empirical sample complexity evaluation. We study how the generalization performance of adversarially robust networks varies with the size of the training dataset. To do so, we train networks with a specific ℓ_∞ adversary while reducing the size of the training set. The training subsets are produced by randomly sub-sampling the complete dataset in a class-balanced fashion. When increasing the number of samples, we ensure that each dataset is a superset of the previous one.

We evaluate the robustness of each trained network to perturbations of varying magnitude (ε_{test}). For each choice of training set size N and fixed attack ε_{test}, we select the best performance achieved across all hyperparameters settings (training perturbations ε_{train} and model size). On all three datasets, we observed that the best standard accuracy is usually achieved for the standard trained network, while the best adversarial accuracy for almost all values of ε_{test} was achieved when training with the largest ε_{train}. We maximize over the hyperparameter settings since we are not interested in the performance of a specific model, but rather in the inherent generalization properties of the dataset independently of the classifier used. Figure 2 shows the results of these experiments.

The plots demonstrate the need for more data to achieve adversarially robust generalization. For any fixed test set accuracy, the number of samples needed is significantly higher for robust generalization. In the SVHN experiments (where we have sufficient training samples to observe plateauing behavior), the standard accuracy reaches its maximum with significantly fewer samples than the adversarial accuracy. We report more details of our experiments in Section H of the supplementary material.

Thresholding experiments. Motivated by our theoretical study of the Bernoulli model, we investigate whether thresholding can also improve the sample complexity of robust generalization against an ℓ_∞ adversary on MNIST.

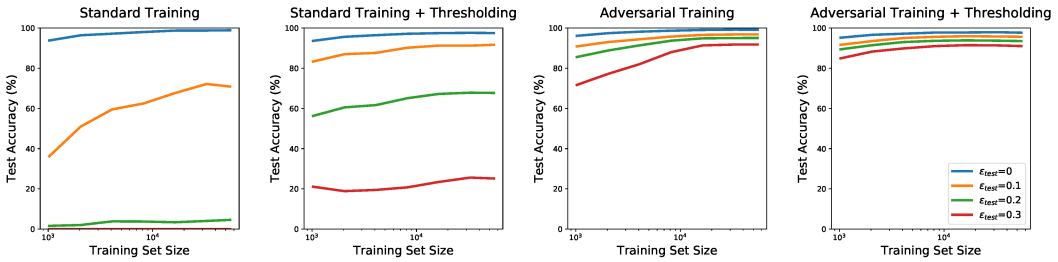

Figure 3: Adversarial robustness to ℓ_∞ attacks on the MNIST dataset for a simple convolution network [36] with and without explicit thresholding filters. For each training set size choice and ε_{test}, we report the best test set accuracy achieved over choice of thresholding filters and ε_{train}. We observe that introducing thresholding filters significantly reduces the number of samples needed to achieve good adversarial generalization.

We repeat the above sample complexity experiments with networks where thresholding filters are explicitly encoded in the model. Here, we replace the first convolutional layer with a fixed thresholding layer consisting of two channels, $\mathrm{ReLU}(x - \varepsilon_{filter})$ and $\mathrm{ReLU}(1 - x - \varepsilon_{filter})$, where x is the input image. Figure 3 shows results for networks trained with such a thresholding layer. For standard trained networks, we use a value of $\varepsilon_{filter} = 0.1$ for the thresholding filters, whereas for adversarially trained networks we set $\varepsilon_{filter} = \varepsilon_{train}$. For each data subset size and test perturbation ε_{test}, we plot the best test accuracy achieved over networks trained with different thresholding filters, i.e., different values of ε. We separately show the effect of explicit thresholding in such networks when they are trained adversarially using PGD.

As predicted by our theory, the networks achieve good adversarially robust generalization with significantly fewer samples when thresholding filters are added. Further, note that adding a simple thresholding layer directly yields nearly state-of-the-art robustness against moderately strong adversaries ($\varepsilon = 0.1$), without any other modifications to the model architecture or training algorithm. It is also worth noting that the thresholding filters could have been learned by the original network architecture, and that this modification only decreases the capacity of the model. Our findings emphasize network architecture as a crucial factor for learning adversarially robust networks from a limited number of samples.

We also experimented with thresholding filters on the CIFAR10 dataset, but did not observe any significant difference from the standard architecture. This agrees with our theoretical understanding that thresholding helps primarily in the case of (approximately) binary datasets.

4 Related Work

Due to the large body of work on adversarial robustness, we focus on related papers that also provide theoretical explanations for adversarial examples. We defer a detailed discussion of related work to Appedix A and discuss here the works most closely related to ours.

Wang et al. [55] study the adversarial robustness of nearest neighbor classifiers. In contrast to our work, the authors give theoretical guarantees for a specific classification algorithm, and do not see a separation in sample complexity between robust and regular generalization. Recent work by Gilmer et al. [20] explores a specific distribution where robust learning is empirically difficult with overparametrized neural networks. The main phenomenon is that even a small natural error rate on their dataset translates to a large adversarial error rate. Our results give a more nuanced picture that involves the sample complexity required for generalization. In our data models, it is possible to achieve an error rate that is essentially zero by using a very small number of samples, whereas the adversarial error rate is still large unless we have seen a lot of samples.

The work of Xu et al. [58] establishes a connection between robust optimization and regularization for linear classification. In particular, they show that robustness to a specific perturbation set is exactly equivalent to the standard support vector machine. Subsequent work by Xu and Mannor [57] builds a deeper connection between robustness and generalization. They prove that for a certain notion of robustness, robust algorithms generalize. Moreover, they show that robustness is a necessary

condition for generalization in an asymptotic sense. Bellet and Habrard [5] gives similar results for metric learning. However, these results do no imply sample complexity bounds since they are asymptotic. Our results stand in stark contrast: we show that generalization can, in simple models, be significantly easier than robustness when sample complexity enters the picture.

Fawzi et al. [18] relate the robustness of linear and non-linear classifiers to adversarial and (semi-) random perturbations. Their work studies the setting where the classifier is fixed and does not encompass the learning task. Fawzi et al. [19] give provable lower bounds for adversarial robustness in models where robust classifiers do not exist. In contrast, we are interested in a setting where robust classifiers exist, but need many samples to learn. Papernot et al. [43] discuss adversarial robustness at the population level. We defer a more detailed discussion of these works to Appendix A.

There is also a long line of work in machine learning on exploring the connection between various notions of margin and generalization, e.g., see [46] and references therein. In this setting, the ℓ_p margin, i.e., how robustly classifiable the data is for ℓ_p^*-bounded classifiers, enables dimension-independent control of the sample complexity. However, the sample complexity in concrete distributional models can often be significantly smaller than what the margin implies.

5 Discussion and Conclusions

The vulnerability of neural networks to adversarial perturbations has recently been a source of much discussion and is still poorly understood. Different works have argued that this vulnerability stems from their discontinuous nature [51], their linear nature [21], or is a result of high-dimensional geometry and independent of the model class [20]. Our work gives a more nuanced picture. We show that for a natural data distribution (the Gaussian model), the model class we train does not matter and a standard linear classifier achieves optimal robustness. However, robustness also strongly depends on properties of the underlying data distribution. For other data models (such as MNIST or the Bernoulli model), our results demonstrate that non-linearities are indispensable to learn from few samples. This dichotomy provides evidence that defenses against adversarial examples need to be tailored to the specific dataset (even for the same type of perturbations) and hence may be more complicated than a single, broad approach. Understanding the interactions between robustness, classifier model, and data distribution from the perspective of generalization is an important direction for future work. We refer the reader to Section B in the appendix for concrete questions in this direction.

The focus of our paper is on adversarial perturbations in a setting where the test distribution (before the adversary's action) is the same as the training distribution. While this is a natural scenario from a security point of view, other setups can be more relevant in different robustness contexts. For instance, we may want a classifier that is robust to small changes between the training and test distribution. This can be formalized as the classification accuracy on *unperturbed* examples coming from an *adversarially* modified distribution. Here, the power of the adversary is limited by how much the test distribution can be modified, and the adversary is not allowed to perturb individual samples coming from the modified test distribution. Interestingly, our lower bound for the Gaussian model also applies to such worst-case distributional shifts. In particular, if the adversary is allowed to shift the mean $\theta^\star$ by a vector in $\mathcal{B}_\infty^\varepsilon$, our proof sketched in Section C transfers to the distribution shift setting. Since the lower bound relies only on a single universal perturbation, this perturbation can also be applied directly to the mean vector.

What do our results mean for robust classification of real images? Our Gaussian lower bound implies that if an algorithm works for all (or most) settings of the unknown parameter $\theta^\star$, then achieving strong ℓ_∞-robustness requires a sample complexity increase that is polynomial in the dimension. There are a few different ways this lower bound could be bypassed. It is conceivable that the noise scale σ is significantly smaller for real image datasets, making robust classification easier. Even if that was not the case, a good algorithm could work for the parameters $\theta^\star$ that correspond to real datasets while not working for most other parameters. To accomplish this, the algorithm would implicitly or explicitly have prior information about the correct $\theta^\star$. While some prior information is already incorporated in the model architectures (e.g., convolutional and pooling layers), the conventional wisdom usually is not to bias the neural network with our priors. Our work suggests that there are trade-offs with robustness here and that adding more prior information could help to learn more robust classifiers.

Acknowledgements

During this research project, Ludwig Schmidt was supported by a Google PhD fellowship and a Microsoft Research fellowship at the Simons Institute for the Theory of Computing. Ludwig was also an intern in the Google Brain team. Shibani Santurkar is supported by the National Science Foundation (NSF) under grants IIS-1447786, IIS-1607189, and CCF-1563880, and the Intel Corporation. Dimitris Tsipras was supported in part by the NSF grant CCF-1553428 and the NSF Frontier grant CNS-1413920. Aleksander Mądry was supported in part by an Alfred P. Sloan Research Fellowship, a Google Research Award, and the NSF grants CCF-1553428 and CNS-1815221.

References

[1] Tensor flow models repository. `https://www.tensorflow.org/tutorials/layers`, 2017.

[2] Anurag Arnab, Ondrej Miksik, and Philip H. S. Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL `http://arxiv.org/abs/1711.09856`.

[3] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. URL `https://arxiv.org/abs/1802.00420`.

[4] Vahid Behzadan and Arslan Munir. Vulnerability of deep reinforcement learning to policy induction attacks. In *International Conference on Machine Learning and Data Mining (MLDM)*, 2017. URL `https://arxiv.org/abs/1701.04143`.

[5] Aurélien Bellet and Amaury Habrard. Robustness and generalization for metric learning. *Neurocomputing*, 2015. URL `https://arxiv.org/abs/1209.1086`.

[6] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton University Press, 2009.

[7] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 2018. URL `https://arxiv.org/abs/1712.03141`.

[8] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[9] Nader H. Bshouty, Nadav Eiron, and Eyal Kushilevitz. PAC learning with nasty noise. In *Algorithmic Learning Theory (ALT)*, 1999. URL `https://link.springer.com/chapter/10.1007/3-540-46769-6_17`.

[10] Nicholas Carlini and David Wagner. Defensive distillation is not robust to adversarial examples. arXiv, 2016.

[11] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, 2016. URL `http://arxiv.org/abs/1608.04644`.

[12] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *Security and Privacy Workshops (SPW)*, 2018. URL `https://arxiv.org/abs/1801.01944`.

[13] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *USENIX Security Symposium*, 2016. URL `https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini`.

[14] Moustapha M Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In *Neural Information Processing Systems (NIPS)*, 2017. URL `https://arxiv.org/abs/1707.05373`.

[15] Ekin D. Cubuk Cubuk, Barret Zoph, Samuel S. Schoenholz, and Quoc V. Le. Intriguing properties of adversarial examples. arXiv, 2017. URL `https://arxiv.org/abs/1711.02846`.

[16] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2004. URL http://doi.acm.org/10.1145/1014052.1014066.

[17] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. arXiv, 2017. URL https://arxiv.org/abs/1712.02779.

[18] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Neural Information Processing Systems (NIPS)*, 2016. URL https://arxiv.org/abs/1608.08967.

[19] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. arXiv, 2018. URL https://arxiv.org/abs/1802.08686.

[20] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. arXiv, 2018. URL https://arxiv.org/abs/1801.02774.

[21] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv, 2014. URL http://arxiv.org/abs/1412.6572.

[22] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick D. McDaniel. Adversarial perturbations against deep neural networks for malware classification. In *European Symposium on Research in Computer Security (ESORICS)*, 2016. URL http://arxiv.org/abs/1606.04435.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL https://arxiv.org/abs/1512.03385.

[24] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defenses: Ensembles of weak defenses are not strong. In *USENIX Workshop on Offensive Technologies*, 2017. URL https://arxiv.org/abs/1706.04701.

[25] Alex Huang, Abdullah Al-Dujaili, Erik Hemberg, and Una-May O'Reilly. Adversarial deep learning for robust detection of binary encoded malware. In *Security and Privacy Workshops (SPW)*, 2018. URL https://arxiv.org/abs/1801.02950.

[26] Sandy H. Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1702.02284.

[27] Peter J. Huber. *Robust Statistics*. Wiley, 1981.

[28] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. URL https://arxiv.org/abs/1707.07328.

[29] Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 1993. URL http://dx.doi.org/10.1137/0222052.

[30] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 1994. URL https://doi.org/10.1023/A:1022615600103.

[31] J Zico Kolter and Eric Wong. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1711.00851.

[32] Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *Security and Privacy Workshops (SPW)*, 2018. URL http://arxiv.org/abs/1702.06832.

[33] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[34] Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. The mnist database of handwritten digits. Website, 1998. URL http://yann.lecun.com/exdb/mnist/.

[35] Daniel Lowd and Christopher Meek. Adversarial learning. In *International Conference on Knowledge Discovery in Data Mining (KDD)*, 2005. URL http://doi.acm.org/10.1145/1081870.1081950.

[36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1706.06083.

[37] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL https://arxiv.org/abs/1511.04599.

[38] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://arxiv.org/abs/1610.08401.

[39] Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial perturbations for deep networks. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. URL http://arxiv.org/abs/1612.06299.

[40] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. URL http://ufldl.stanford.edu/housenumbers/.

[41] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Symposium on Security and Privacy (SP)*, 2016. URL https://arxiv.org/abs/1511.04508.

[42] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *European Symposium on Security and Privacy (EuroS&P)*, 2016. URL https://arxiv.org/abs/1511.07528.

[43] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. In *European Symposium on Security and Privacy (EuroS&P)*, 2018. URL https://arxiv.org/abs/1611.03814.

[44] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1801.09344.

[45] Phillippe Rigollet and Jan-Christian Hütter. High-dimensional statistics. *Lecture notes*, 2017. URL http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf.

[46] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. URL http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/.

[47] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Conference on Computer and Communications Security (CCS)*, 2016. URL http://doi.acm.org/10.1145/2976749.2978392.

[48] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1710.10571.

[49] Liwei Song and Prateek Mittal. Inaudible voice commands. In *Conference on Computer and Communications Security (CCS)*, 2017. URL http://arxiv.org/abs/1708.07238.

[50] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. arXiv, 2017. URL http://arxiv.org/abs/1710.08864.

[51] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. URL http://arxiv.org/abs/1312.6199.

[52] Florian Tramèr, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. The space of transferable adversarial examples. arXiv, 2017. URL http://arxiv.org/abs/1704.03453.

[53] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018. URL http://arxiv.org/abs/1705.07204.

[54] Abraham Wald. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, 1945.

[55] Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. Analyzing the robustness of nearest neighbors to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. URL http://proceedings.mlr.press/v80/wang18c/wang18c.pdf.

[56] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1801.02612.

[57] Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 2012. URL https://arxiv.org/abs/1005.2243.

[58] Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research (JMLR)*, 2009. URL http://www.jmlr.org/papers/v10/xu09b.html.

[59] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed System Security Symposium (NDSS)*, 2017. URL https://arxiv.org/abs/1704.01155.

[60] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darell, and Dawn Song. Can you fool AI with adversarial examples on a visual Turing test? arXiv, 2017. URL http://arxiv.org/abs/1709.08693.

[61] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016. URL http://arxiv.org/abs/1605.07146.

[62] Guoming Zhang, Chen Yan, Xiaoyu Ji, Taimin Zhang, Tianchen Zhang, and Wenyuan Xu. Dolphinatack: Inaudible voice commands. In *Conference on Computer and Communications Security (CCS)*, 2017. URL http://arxiv.org/abs/1708.09537.