Query Complexity of Adversarial Attacks

Grzegorz Głuch [1] Rüdiger Urbanke [1]

Abstract

There are two main attack models considered in the adversarial robustness literature: black-box and white-box. We consider these threat models as two ends of a fine-grained spectrum, indexed by the number of queries the adversary can ask. Using this point of view we investigate how many queries the adversary needs to make to design an attack that is comparable to the best possible attack in the white-box model. We give a lower bound on that number of queries in terms of entropy of decision boundaries of the classifier. Using this result we analyze two classical learning algorithms on two synthetic tasks for which we prove meaningful security guarantees. The obtained bounds suggest that some learning algorithms are inherently more robust against query-bounded adversaries than others.

1. Introduction

Modern neural networks achieve high accuracy on tasks such as image classification (Krizhevsky et al., 2012) or speech recognition (Collobert & Weston, 2008). However, they are typically susceptible to small, adversarially-chosen perturbations of the inputs (Szegedy et al., 2014; Nguyen et al., 2015): more precisely, given a correctly-classified input x, one can typically find a small perturbation δ such that $x+\delta$ is misclassified by the network while to the human eye this perturbation is not perceptible.

There are two main threat models considered in the literature: black-box and white-box. In the white-box model, on the one hand, the attacker (Biggio et al., 2013; Zheng et al., 2019) is assumed to have access to a full description of the model. For the case of neural networks that amounts to a knowledge of the architecture and the weights. In the black-box model, on the other hand, the adversary (Papernot et al., 2017; Chen et al., 2017; Liu et al., 2016; Xiao et al., 2018; Hayes & Danezis, 2017) can only observe the input-output behavior of the model. Many defenses have been proposed to date. To mention just some – adversarial learning (Goodfellow et al., 2014; Madry et al., 2018; Tramèr et al., 2018; Xiao et al., 2019), input denoising (Bhagoji et al., 2017; Xu et al., 2017), or more recently, randomized smoothing (Lécuyer et al., 2019; Cohen et al., 2019; Salman et al., 2019; Gluch & Urbanke, 2019). Unfortunately, most heuristic defenses break in the presence of suitably strong adversaries (Carlini & Wagner, 2017; Athalye et al., 2018; Uesato et al., 2018) and provable defenses are often impractical or allow only very small perturbations. Thus a full defense remains elusive. The current literature on this topic is considerable. We refer the reader to Chakraborty et al. (2018) for an overview of both attacks and defenses and to Bhambri et al. (2019) for a survey focused on black-box attacks.

We consider black-box and white-box models as the extreme points of a spectrum parameterized by the number of queries allowed for the adversary. This point of view is related to Ilyas et al. (2018) where the authors design a black-box attack with a limited number of queries. Intuitively, the more queries the adversary can make the more knowledge he gains about the classifier. When the number of queries approaches infinity then we transition from a black-box to a white-box model as in this limit the adversary knows the classifying function exactly. Using this point of view we ask:

How many queries does the adversary need to make to reliably find adversarial examples?

By “reliably” we mean comparable with the information-theoretic white-box performance. To be more formal, we assume that there is a distribution $\mathcal{D}$ and a high-accuracy classifier f that maps $\mathbb{R}^d$ to classes $\mathcal{Y}$. The adversary $\mathcal{A}$ only has black-box access to f. Moreover, $\epsilon \in \mathbb{R}_+$ is an upper bound on the norm (usually ℓ_p norm) of the allowed adversarial perturbation. Assume that f is susceptible to ϵ-bounded adversarial perturbations for an η-fraction of the underlying distribution $\mathcal{D}$. The quantity η is the largest error an adversary, who has access to unbounded computational resources and fully knows f, can achieve. We ask: How many queries to the classifier f does $\mathcal{A}$ need to make in order to be able to find adversarial examples for say an

[1] School of Computer and Communication Sciences, EPFL, Switzerland. Correspondence to: Grzegorz Głuch <grzegorz.gluch@epfl.ch>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

$\eta/2$-fraction of the distribution $\mathcal{D}$? This question is similar to problems considered in Ashtiani et al. (2020). The difference is that in Ashtiani et al. (2020) the authors define the query complexity of the adversary as a function of the number of points for which the adversarial examples are to be found. Moreover, they require the adversary to be perfect, that is to find adversarial examples whenever they exist. This stands in contrast to our approach that only requires the adversary to succeed for say a 1/2 fraction of the adversarial examples. The question we ask is also similar to ideas in Cullina et al. (2018). In this paper the authors consider a generalization of PAC learning and ask how many queries an algorithm requires in order to learn robustly. Similar questions were also asked in Yin et al. (2019) and (Schmidt et al., 2018). The difference is that we focus on the query complexity of the attacker and not the defender. Diochnos et al. (2018); Gourdeau et al. (2019) discuss the use of membership queries as part of converting a black-box classifier to a white-box classifier.

Our contributions. We introduce a new notion - the **query complexity** (QC) of adversarial attacks. This notion unifies the two most popular attack models and enables a systematic study of robustness of learning algorithms against query-bounded adversaries.

Our findings are the following: the higher the entropy of the decision boundaries that are created by the learning algorithm the more secure is the resulting system in our attack model. We first prove a general lower bound on the QC in terms of the entropy of decision boundaries. Then, using this result, we present two scenarios for which we are able to show meaningful lower bounds on the QC. The first one is a simple 2-dimensional distribution and a nearest neighbor algorithm. For this setting we are able to prove a strong query lower bound of $\Theta(m)$, where m is the number of samples on which the classifier was "trained". For the second example we consider the well-known adversarial spheres distribution, introduced in the seminal paper Gilmer et al. (2018). For this learning task we argue that quadratic neural networks have a query lower bound of $\Theta(d)$, where d is the dimensionality of the data. We discuss why certain learning algorithms like linear classifiers and also neural networks might be far less secure than KNN against query-bounded adversaries. Finally, we use the lower bound on the QC in terms of entropy to prove, for a broad class of learning algorithms, a security guarantee against query-bounded adversaries that grows with accuracy.

There exist tasks for which it is easy to find high-accuracy classifiers but finding robust models is infeasible. E.g., in Bubeck et al. (2019) the authors describe a situation where it is information-theoretically easy to learn robustly but there is no algorithm in the statistical query model that computes a robust classifier. In Bubeck et al. (2018) an even stronger result is proven. It is shown that under a standard cryptographic assumption there exist learning tasks for which no algorithm can efficiently learn a robust classifier. Finally, in Tsipras et al. (2019) it was shown that robust and accurate classifiers might not even exist. The query-bounded point of view shows a way to address these fundamental difficulties – even for tasks for which it is impossible to produce a model that is secure against a resource-unbounded adversary, it might be possible to defend against a query-bounded adversary.

Organization of the paper. In Section 2 we formally define the threat model and the query complexity of adversarial attacks. In Section 7 we show that a security guarantee against query-bounded adversaries that grows with accuracy for a rich class of learning algorithms. In Section 4 and 5 we analyze the query complexity of KNN and Quadratic Neural Network learning algorithms respectively. In Section 6 we present a universal defense against query-bounded adversaries. Finally, in Section 9 we summarize the results and discuss future directions. We defer most of the proofs to the appendix.

2. The Query Complexity of adversarial attacks

We start by formally defining the *threat model*. For a *data distribution* $\mathcal{D}$ over $\mathbb{R}^d$ and a set $A \subseteq \mathbb{R}^d$ let $\mu(A) := \mathbb{P}_{X \sim \mathcal{D}}[X \in A]$. For simplicity we consider only *separable* binary classification tasks. Such tasks are fully specified by $\mathcal{D}$ as well as a *ground truth* $h : \mathbb{R}^d \to \{-1, 1\}$. For a binary classification task with a ground truth $h : \mathbb{R}^d \to \{-1, 1\}$ and a classifier $f : \mathbb{R}^d \to \{-1, 1\}$ we define the *error set* as $E(f) := \{x \in \mathbb{R}^d : f(x) \neq h(x)\}$. Note that with this definition it might happen that $E(f) \not\subseteq \text{supp}(\mathcal{D})$. For $x \in \mathbb{R}^d$ and $\epsilon > 0$ we write $B_\epsilon(x)$ to denote the *closed ball* with center x and radius ϵ and B_ϵ to denote the *closed ball* with center 0 and radius ϵ. We say that a function $p : \mathbb{R}^d \to \mathbb{R}^d$ is an ϵ*-perturbation* if for all $x \in \mathbb{R}^d$ we have $\|p(x) - x\|_2 \leq \epsilon$. For $n \in \mathbb{N}$ we denote the set $\{1, 2, \ldots, n\}$ by $[n]$. For $x, y \in \mathbb{R}^d$ we will use $[x, y]$ to denote the closed line segment between x and y. For $A, B \subseteq \mathbb{R}^d$ we define $A + B := \{x + y : x \in A, y \in B\}$. We use m to denote the sample size.

Definition 1 (**Risk**). *Consider a separable, binary classification task with a ground truth $h : \mathbb{R}^d \to \{-1, 1\}$. For a classifier $f : \mathbb{R}^d \to \{-1, 1\}$ we define the **R**isk as* $R(f) := \mathbb{P}_X[f(X) \neq h(X)]$.

Definition 2 (**Adversarial risk**). *Consider a binary classification task with separable classes with a ground truth $h : \mathbb{R}^d \to \{-1, 1\}$. For a classifier $f : \mathbb{R}^d \to \{-1, 1\}$ and $\epsilon \in \mathbb{R}_{\geq 0}$ we define the **A**dversarial **R**isk as:*

$$AR(f, \epsilon) := \mathbb{P}_X[\exists\, \gamma \in B_\epsilon : f(X + \gamma) \neq h(X + \gamma)].$$

an ϵ-perturbation p we define:

$$AR(f,p) := \mathbb{P}_X[f(p(X)) \neq h(p(X))],$$

to be the adversarial risk of a specific perturbation function p.

Note: In the literature other definitions were also considered, see Diochnos et al. (2018); Gourdeau et al. (2019). In order to keep the exposition simple, we restrict our discussion to our definition of adversarial risk and to ℓ_2-bounded adversarial perturbations. Other norms can of course be considered and might be important in practice.

Definition 3 (Query-bounded adversary). *For $\epsilon \in \mathbb{R}_{\geq 0}$ and $f : \mathbb{R}^d \to \{-1,1\}$ a q-bounded adversary with parameter ϵ is a deterministic algorithm*[*] *$\mathcal{A}$ that asks at most $q \in \mathbb{N}$ (potentially adaptive) queries of the form $f(x) \overset{?}{=} 1$ and outputs an ϵ-perturbation $\mathcal{A}(f) : \mathbb{R}^d \to \mathbb{R}^d$.*

Definition 4 (Query complexity of adversarial attacks). *Consider a binary classification task T for separable classes with a ground truth $h : \mathbb{R}^d \to \{-1,1\}$ and a distribution $\mathcal{D}$. Assume that there is a learning algorithm ALG for this task that given $S \sim \mathcal{D}^m$ learns a classifier $ALG(S) : \mathbb{R}^d \to \{-1,1\}$. For $\epsilon \in \mathbb{R}_{\geq 0}$ define the Query Complexity of adversarial attacks on ALG with respect to (T,m,ϵ) and denote it by $QC(ALG,T,m,\epsilon)$: It is the minimum $q \in \mathbb{N}$ so that there exists a q-bounded adversary $\mathcal{A}$ with parameter ϵ such that $\mathbb{P}_{S\sim\mathcal{D}^m}$ of the event*

$$AR(ALG(S), \mathcal{A}(ALG(S))) \geq \frac{1}{2} AR(ALG(S), \epsilon)$$

is at least 0.99.

In words, it is the minimum number of queries that is needed so that there exists an adversary who can achieve an error of half the maximum achievable error (with high probability over the data samples). Note that it follows from Definitions 3 and 4 that $\mathcal{A}$ is computationally unbounded, knows the distribution $\mathcal{D}$ and the ground truth h of the learning task and also knows the learning algorithm ALG. The only restriction that is imposed on $\mathcal{A}$ is the number of allowed queries. What is important is that $\mathcal{A}$ does *not* know S nor the potential randomness of ALG (in the generalized setting ALG can be randomized, see Definition 5) – this is what makes the QC non-degenerate. To see this, observe that if $\mathcal{A}$ knew S and ALG was deterministic then $\mathcal{A}$ could achieve $AR(\text{ALG}(S), \mathcal{A}(\text{ALG}(S))) = AR(\text{ALG}(S), \epsilon)$ without asking any queries. This is because $\mathcal{A}$ can for every point x check if there exists $\gamma \in B_\epsilon$ such that $\text{ALG}(S)(X+\gamma) \neq h(X)$, as $\mathcal{A}$ can compute $\text{ALG}(S)$ without asking any queries. This allows $\mathcal{A}$ to achieve adversarial risk of $AR(\text{ALG}(S), \epsilon)$ (see Definition 3).

Defnition 4 was guided by experiments. For instance, in Papernot et al. (2017), in order to attack a neural network f the adversary trains a new neural network $\hat{f}$. This $\hat{f}$ acts as an approximation of f. She does so by creating a training set, which is labeled using f and then training $\hat{f}$ on this dataset. The concept of transferability allows the adversary to ensure that f will also misclassify inputs misclassified by $\hat{f}$. This phenomenon is reflected in our definition – after the initial phase of querying f (and training $\hat{f}$) the adversary no longer has access to f. This means that after this phase the adversary implicitly constructed an ϵ-perturbation p as in Definition 2. One can also consider different definitions. For instance ask about the number of queries required to attack a *given point* or measure the adversarial risk in *absolute terms* (instead of 1/2 of the white-box performance). Both of these questions are valid and are of interest in their own right. For the sake of definiteness we choose what we consider to be one important viewpoint, a viewpoint that is well motivated by experiments.

Definition 4 can be generalized to incorporate randomness in the learning algorithm. Intuitively, the randomness in ALG can increase the entropy of the learning process and that in turn may lead to a higher QC. Further, both the approximation constant (which is chosen to be $1/2$ in Definition 4) as well as the success probability can also be generalized.

Definition 5 (Query complexity of adversarial attacks - generalized). *Consider a binary classification task T for separable classes with a ground truth $h : \mathbb{R}^d \to \{-1,1\}$ and a distribution $\mathcal{D}$. Assume that there is a* ***randomized*** *learning algorithm ALG for this task that given $S \sim \mathcal{D}^m$ and a sequence of random bits $B \sim \mathcal{B}$ learns a classifier $ALG(S,B) : \mathbb{R}^d \to \{-1,1\}$. For $\epsilon \in \mathbb{R}_{\geq 0}, \kappa, \alpha \in [0,1]$ define the Query Complexity of the adversarial attacks on ALG with respect to $(T,m,\epsilon,\alpha,\kappa)$ and denote it by $QC(ALG,T,m,\epsilon,\alpha,\kappa)$: It is the minimum $q \in \mathbb{N}$ such that there exists a q-bounded adversary $\mathcal{A}$ with parameter ϵ such that $\mathbb{P}_{S\sim\mathcal{D}^m, B\sim\mathcal{B}}$ of the event*

$$AR(ALG(S,B), \mathcal{A}(ALG(S,B))) \geq \alpha AR(ALG(S,B), \epsilon)$$

is at least $1-\kappa$.

If the above holds for $\mathcal{A}$ we will refer to α as the ***approximation constant*** *of $\mathcal{A}$ and to κ as the* ***error probability*** *of $\mathcal{A}$.*

For the sake of clarity whenever possible we will restrict ourselves for the most part to Definition 4. This eliminates two parameters from our expressions and restricts the attention to deterministic learning algorithms. Only when the

[*] We use *algorithm* here since this seems more natural. But we do not limit the attacker computationally nor are we concerned with questions of computability. Hence, *function* would be equally correct.

distinction becomes important will we refer to Definition 5.

Summary: The query complexity of adversarial attacks is the minimum q for which there exists a q-bounded adversary that carries out a successful attack. Such adversaries are computationally unbounded, know the learning task and the learning algorithm but *don't* know the training set.

3. High-entropy decision boundaries lead to robustness

The decision boundary of a learning algorithm applied to a given task can be viewed as the outcome of a random process: (i) generate a training set and, (ii) apply to it the, potentially randomized, learning algorithm. Recall, see Definitions 4 and 5, that a query-bounded adversary does not know the sample on which the model was trained nor the randomness used by the learner. This means that if the decision boundary has high entropy then the adversary needs to ask many questions to recover the boundary to a high degree of precision. This suggest that high-entropy decision boundaries are robust against query-bounded adversaries since intuitively it is clear that an approximate knowledge of the decision boundary is a prerequisite for a successful attack.

Next we present a result that makes this intuition formal. Before delving into the details let us explain the intuition behind this approach. Let us recall the set-up. The classifier is trained on a sample S that is unknown to the adversary. This classifier has a particular error set. We say that an adversary succeeds if, after asking some queries, it manages to produce an ϵ-perturbation with the property that this perturbation moves "sufficient" mass into the error set of the classifier. Here, sufficient means at least half of what is possible if the adversary had known the classifier exactly. Let us say in this case that an ϵ-perturbation is *consistent* with an error set.

The following theorem states that if for every ϵ-perturbation the probability that an error set of a classifier is consistent with that perturbation is small then the QC is high. This is true since if for every ϵ-perturbation only a small fraction of probability space (i.e., the possible classifiers) is consistent with this perturbation then $\mathcal{A}$'s protocol has to return many distinct ϵ-perturbations depending on the outcome of its queries. And to distinguish which perturbation it should return it has to ask many queries.

Theorem 1. ***[Reduction.]*** *Let $\epsilon \in \mathbb{R}_{\geq 0}$ and let T be a binary classification task on $\mathbb{R}^d$ with separable classes. Let ALG be a randomized learning algorithm for T that uses m samples. Then for every $\kappa \in [0,1]$ the following holds:*

$$\begin{aligned} &QC(ALG, T, m, \epsilon, 1/2, \kappa) \\ &\geq \log\left(\frac{1-\kappa}{\sup_{p:\ \epsilon\text{-perturbation}} \mathbb{P}_{S\sim\mathcal{D}^m, B\sim\mathcal{B}}\left[\mathcal{E}(S,B,p)\right]}\right), \end{aligned}$$

where the event $\mathcal{E}(S,B,p)$ is defined as:

$$\mu(p^{-1}(E(ALG(S,B)))) \geq \frac{AR(ALG(S,B),\epsilon)}{2}.$$

Remark 1. *For the sake of clarity and consistency with the standard setup we fixed the approximation constant to be equal $1/2$. We note however, that Theorem 1 (and its proof) is also true for all approximation constants.*

Summary: Theorem 1 is a key ingredient in most of our results. It gives a lower bound on the QC in terms of a geometric-like notion of entropy of error sets. This is often much easier to compute than to analyze the inner workings of a particular learning algorithm.

4. Entropy due to locality – K-NN algorithms

Let us now analyze the QC of K-Nearest Neighbor (K-NN) algorithms. Nearest neighbor algorithms are among the simplest and most studied algorithms in machine learning. They are also widely used as a benchmark. It was shown in Cover & Hart (1967) that for a sufficiently large training set, the risk of the 1-NN learning rule is upper bounded by twice the optimal risk. It is also known that these methods suffer from the "curse of dimensionality" – for d dimensional distributions they typically require $m = 2^{\Theta(d\log(d))}$ many samples. That is why in practice one often uses some dimensionality reduction subroutine before applying K-NN. Moreover, K-NN techniques are one of the few learning algorithms that do not require any learning. In the naive implementation all computation is deferred until function evaluation. This is related to the most interesting fact from our perspective, namely that **the classification rule of the K-NN algorithm depends only on the local structure of the training set**.

We argue that this property makes K-NN algorithms secure against query-bounded adversaries. Intuitively, if the adversary $\mathcal{A}$ wants to achieve a high adversarial risk she needs to understand the global structure of the decision boundary. But if the classification rule is only very weakly correlated between distant regions of the space then this intuition suggests that $\mathcal{A}$ may need to ask $\Theta(m)$ many queries to guarantee a high adversarial risk. This is consistent with the entropy point of view. Moreover there are experimental results (see Wang et al. (2018); Papernot et al. (2016)) that show that it is hard to attack K-NN classifiers in the black-box model.

We make these intuitions formal in the following sense. We design a synthetic binary learning task in $\mathbb{R}^2$, where the data is uniformly distributed on two parallel intervals – correspondiing to the two classes. We then show a $\Theta(m)$ lower bound for the QC of 1-NN algorithm for this learning task. This means that the number of queries the adversary needs to make to attack 1-NN is proportional to the number of samples on which the algorithm was "trained". We conjecture that a similar behavior occurs in higher dimensions as well.

4.1. K-NN – QC lower bounds

Consider the following distribution. Let $m \in \mathbb{N}$ and $z \in \mathbb{R}_+$. Let $L_-, L_+ \subseteq \mathbb{R}^2$ be two parallel intervals of length m placed at distance z apart. More formally, $L_- := [(0,0),(m,0)]$, $L_+ = [(0,z),(m,z)]$. Let the binary learning task $T_{\text{intervals}}(z)$ be as follows. We generate $\bar{x} \in \mathbb{R}^2$ uniformly at random from the union $L_- \cup L_+$. We assign the label $y = -1$ if $\bar{x} \in L_-$ and $y = +1$ otherwise. In Figure 1 we visualize a decision boundary of the 1-NN algorithm for a random $S \sim \mathcal{D}^m$ on the $T_{\text{intervals}}$. Horizontal lines, black and gray, represent the two classes, crosses are data points, white and gray regions depict the classification rule and the union of red intervals is equal to the error set. We also include more visualizations in the appendix. The main result of this subsection is:

Theorem 2. *There exists a function $\lambda : \mathbb{R}^+ \to (0,1)$ such that the 1-Nearest Neighbor (1-NN) algorithm applied to the learning task $T_{intervals}(z)$ satisfies:*

$$QC(\text{1-NN}, T_{intervals}(z), 2m, z/10, 1 - \lambda(z), 0.1) \geq \Theta(m),$$

provided that $z = \Omega(1)$.

Summary: The K-NN algorithm learns classification rules that depend only on the local structure of the data. This implies high-entropy decision boundaries, which in turn leads to robustness against query-bounded adversaries. The QC of 1-NN scales at least linearly with the size of the training set.

5. Entropy due to symmetry - Quadratic Neural Networks

In this section we analyze the QC of Quadratic Neural Networks (QNN) applied to a learning task defined in Gilmer et al. (2018). Let $S_r^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = r\}$. The distribution $\mathcal{D}$ is defined by the following process: generate $x \sim U[S_1^{d-1}]$ and $b \sim U\{-1,1\}$ (where U denotes the uniform distribution). If $b = -1$ return $(x, -1)$, otherwise return $(1.3x, +1)$. The associated ground truth is defined as $h(x) = -1$ for $x \in \mathbb{R}^d, \|x\|_2 \leq 1.15$ and $h(x) = 1$ otherwise .

The QNN is a single hidden-layer network where the activation function is the quadratic function $\sigma(x) = x^2$. There are no bias terms in the hidden layer. The output node computes the sum of the activations from the hidden layer, multiplies them by a scalar and adds a bias. If we assume that the hidden layer contains h nodes then the network has $d \cdot h + 2$ parameters. It was shown in Gilmer et al. (2018) that the function that is learned by QNN has the form $y(x) = \sum_{i=1}^d \alpha_i z_i^2 - 1$, where the α_i's are scalars that depend on the parameters of the network and $z = M(x)$ for some orthogonal matrix M. The decision boundary is thus $\sum_{i=1}^d \alpha_i z_i^2 = 1$, which means that it is an ellipsoid centered at the origin.

In a series of experiments performed for the Concentric Spheres (CS) dataset in Gilmer et al. (2018) it was shown that a QNN trained with $N = 10^6$ many samples with $d = 500$ and $h = 1000$ learns a classifier with an estimated error of approximately 10^{-20} but the adversarial risk η is high and is estimated to be $1/2$ when $\epsilon \approx 0.18$. On the theoretical side, it was proven in Gluch & Urbanke (2019) (see Section 9.1) that

$$\epsilon \leq O\left(\frac{\log(\eta/\delta)}{d}\right). \tag{1}$$

In words, (1) gives an upper bound on the biggest allowed perturbation ϵ in terms of the risk δ, the adversarial risk η and the dimension d. In particular if we want the classifier to be adversarially robust for $\epsilon = \Theta(1)$ (that is for perturbations comparable with the separation between the two classes) then $\delta = 2^{-\Omega(d)}$. Even robustness of only $\epsilon = \Theta(1/\sqrt{d})$ requires the risk to be as small as $\delta = 2^{-\Omega(\sqrt{d})}$. These results paint a bleak picture of the adversarial robustness for CS.

The QC point-of-view is more optimistic. Using results from Section 7 we first show that if the network learns classifiers with risk $2^{-\Omega(k)}$ then it automatically leads to a lower bound on the QC of $\Theta(k)$. Moreover, for a simplified model of the network, we show that even if the risk of the learned classifier is only a small constant, say 0.01, then this results in a lower bound on the QC of $\Theta(d)$ for perturbations of $\Theta(1/\sqrt{d})$. Using (1) our result guarantees security against $\Theta(d)$-bounded adversaries for perturbations which are $\Theta(\sqrt{d})$ times bigger than the best possible against unbounded adversaries. This shows that restricting the power of the adversary can make a significant difference.

We argue that the obtained $\Theta(d)$ lower bound is close to the real QC for this algorithm and learning task. Observe that the decision boundary of the network is an ellipsoid which can be described by $O(d^2)$ parameters (d^2 for the rotation matrix and d for lengths of principal axes). This suggest that it should be possible to design a $O(d^2)$-bounded adversary that succeeds on this task. Assuming that this is indeed the

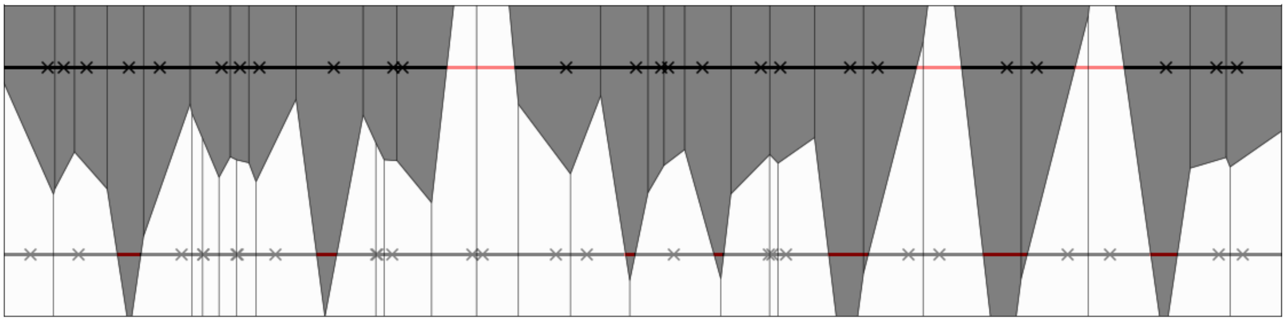

Figure 1. A random decision boundary of 1-NN for $T_{\text{intervals}}$.

case, our lower bound is only a factor $O(d)$ away from the optimum.

The results of this section can be understood in the following way. The relatively simple structure of the decision boundaries allows the adversary to attack the model with only $O(d^2)$ queries. There is however enough entropy in the network to guarantee a lower bound for the QC of $\Theta(d)$. This entropy intuitively comes from the rotational invariance of the dataset and in turn of the learned decision boundary. We conjecture that algorithms like linear classifiers (e.g., SVMs) exhibit a similar behavior. That is, for natural learning tasks they are robust against q-bounded adversaries only for $q = O(\text{poly}(d))$. The reason is that all these algorithms generate classifiers with relatively simple decision boundaries which can be described by $O(\text{poly}(d))$ parameters.

But this is not the end of the story for CS. Our results don't preclude the possibility that there exist a learning algorithm that is secure against q-bounded adversaries for $q \gg d$. In fact in Section 6 we present an off-the-shelf solution that can be applied to CS dataset and which, by injecting entropy, achieves security against k-bounded adversaries for $\epsilon = \Theta\left(\frac{1}{k\sqrt{d}}\right)$.

5.1. Quadratic Neural Networks – QC lower bounds

Using the results from Section 7 one can show that increased accuracy leads to increased robustness. More precisely if QNN has a risk of $2^{-\Omega(k)}$ then it is secure against $\Theta(k)$-bounded adversaries for $\epsilon = \Theta(1)$. The proof of this fact is deferred to the appendix.

Now we argue that also in the case where the risk achieved by the network is as large as a constant then QNN are still robust against $\Theta(d)$-bounded adversaries. We first argue that any reasonable optimization algorithm applied to QNN for the CS learning task gives rise to a distribution on error sets that is rotational invariant. This follows from the fact that $\mathcal{D}$ itself is rotational invariant. Now observe that for QNN the error sets are of the form: $\left\{x \in \mathbb{R}^d : \|x\|_2 \leq 1.15, \sum_{i=1}^d \alpha_i z_i^2 > 1\right\} \cup \left\{x \in \mathbb{R}^d : \|x\|_2 > 1.15, \sum_{i=1}^d \alpha_i z_i^2 < 1\right\}$, as the decision boundary learned by QNN is defined by $\sum_{i=1}^d \alpha_i z_i^2 = 1$, where $z = Mx$ for some orthonormal matrix M. These sets might be quite complicated as they are basically defined as the set difference of a ball and an ellipsoid. We will refer to the real distribution on error sets of QNN as $\mathcal{E}_{\text{QNN}}$. We assume that the standard risk of classifiers learned by the QNN is concentrated around a constant δ.

Intuitively a "complicated" (high entropy) distribution on error sets results in a high QC and a "simple" (low entropy) distribution results in a low QC. In the rest of this section we first introduce a set of "simple", artificial distributions over error sets and then we state QC lower bounds for these distributions. Formal definitions are presented in Definition *Distributions on Spherical Caps* in the appendix, here we give a short description of what they are. For $y \in S_1^{d-1}$ let $\text{cap}(y, r, \tau) := B_r \cap \{x \in \mathbb{R}^d : \langle x, y\rangle \geq \tau\}$. Let $\tau : [0,1] \to [0,1]$ be such that for every $\delta \in [0,1]$ we have $\nu(\text{cap}(\cdot, 1, \tau(\delta)))/\nu(S_1^{d-1}) = \delta$, where ν is a $d-1$ dimensional measure on the sphere S_1^{d-1}. For $k \in \mathbb{N}_+$ let:

$$E_-(k) := \text{cap}(e_1, 1.15, \tau(\delta/k)) \setminus B_{1.15/1.3}$$

$$E_+(k) := \text{cap}(e_1, 1.495, 1.3\tau(\delta/k)) \setminus B_{1.15},$$

where e_1 is a standard basis vector. Note that for every k we have $1.3 \cdot E_-(k) = E_+(k)$. For $\delta \in (0,1), k \in \mathbb{N}_+$ the distributions are: $\text{Cap}(\delta)$ - randomly rotated either $E_-(1)$ or $E_+(1)$, chosen uniformly at random; $\text{Cap}_k^{i.i.d}(\delta)$ - union of k i.i.d. randomly rotated sets each either $E_-(k)$ or $E_+(k)$, chosen uniformly at random; $\text{Cap}_k^{\mathcal{G}}(\delta)$ - k randomly rotated sets, each either $E_-(k)$ or $E_+(k)$, chosen uniformly at random; the relative positions of cap's normal vectors are determined by $\mathcal{G}$, where $\mathcal{G}$ is a given distribution on $(S_1^{d-1})^k$.

We conjecture that $\text{Cap}(0.01)$, $\text{Caps}_d^{\text{i.i.d.}}(0.01)$, $\text{Caps}_d^{\mathcal{G}}(0.01)$ have QCs that are no larger than the QC of $\mathcal{E}_{\text{QNN}}$ that

Table 1. QC for CS

Error distribution	Lower bound
$\text{Cap}(0.01)$	$\Theta(d)$
$\text{Caps}_k^{\text{i.i.d.}}(0.01)$	$\Theta(d)$[†]
$\text{Caps}_k^{\mathcal{G}}(0.01)$	$\Theta(d/k)$[†]

achieves standard risk 0.01. The intuitive reason is that they contain less entropy than $\mathcal{E}_{\text{QNN}}$ and so it should be easier to attack these distributions. In Lemma 1 we prove a $\Theta(d)$ lower bound for $\text{Cap}(0.01)$ and, in the appendix, we give a matching upper bound of $\Theta(d)$. Also in the appendix, we give two reductions that lower-bound QC of $\text{Caps}_d^{\text{i.i.d.}}$ and $\text{Caps}_d^{\mathcal{G}}$ based on a conjecture (see *Cap conjecture* in the appendix). We summarize the proved and conjectured lower bounds in Table 1.

Lemma 1 (Lower bound for Cap). *There exists $\lambda > 0$ such that if a q-bounded adversary $\mathcal{A}$ succeeds on Cap(0.01) with approximation constant $\geq 1 - \lambda$, error probability $2/3$ for $\epsilon = \tau(0.01)$. Then*

$$q \geq \Theta(d).$$

Summary: Quadratic neural networks have simple decision boundaries - they are of the form of ellipsoids. But due to the rotational symmetry there is sufficient entropy to guarantee robustness against $\Theta(d)$-bounded adversaries. [‡]

6. How to increase the entropy of an existing scheme – a universal defense

It was proven in Gluch & Urbanke (2019) that there exists a universal defense against adversarial attacks. The defense algorithm gets as an input access to a high accuracy classifier f and outputs a new classifier g that is adversarially robust. The idea of the defense is based on randomized smoothing (Cohen et al., 2019; Salman et al., 2019) and random partitions of metric spaces. Simple rephrasing of Theorem 5 from Gluch & Urbanke (2019) in the language of the QC of adversarial attacks gives the following:

Theorem 3. *For every $d \in \mathbb{N}_+$ there exists a randomized algorithm DEF satisfying the following. It is given as input access to an initial classifier $\mathbb{R}^d \to \{-1, 1\}$ and provides oracle access to a new classifier $\mathbb{R}^d \to \{-1, 1\}$. For every separable binary classification task T in $\mathbb{R}^d$ with separation ϵ the following conditions hold. Let ALG be a learning algorithm for T that uses m samples. Then for every $S \sim \mathcal{D}^m$ we have $R(\mathit{DEF}(\mathit{ALG}(S))) \leq 2R(\mathit{ALG}(S))$ and for every $\epsilon' > 0$:*

$$QC(\mathit{DEF} \circ \mathit{ALG}, T, m, \epsilon') \geq \Theta\left(\frac{\epsilon}{\sqrt{d} \cdot \epsilon'}\right).$$

Summary: There exists a universal defense that can be applied on top of any learning algorithm to make it secure against query-bounded adversaries. Roughly speaking, it works by injecting additional randomness to increase the entropy of the final classifier.

7. Robustness and accuracy – foes no more

It was argued in Tsipras et al. (2019) that there might be an inherent tension between accuracy and adversarial robustness. We argue that this potential tension disappears for a rich class of learning algorithms if we consider q-bounded adversaries. We show that if a learning algorithm satisfies a particular natural property then there is a lower bound for the QC of this algorithm that *grows* with accuracy.

Theorem 4. *For every $\epsilon \in \mathbb{R}_{\geq 0}, C, \delta, \eta \in \mathbb{R}_+$ and T a binary classification task on $\mathbb{R}^d$ with separable classes the following conditions hold. If ALG is a learning algorithm for T and satisfies the following properties:*

1. *$\forall x \in supp(\mathcal{D}) + B_\epsilon$, $\mathbb{P}_{S \sim \mathcal{D}^m}[\mathit{ALG}(S)(x) \neq h(x)] \leq C \cdot \delta$,*
2. *$\mathbb{P}_{S \sim \mathcal{D}^m}[AR(\mathit{ALG}(S), \epsilon) \geq \eta] \geq 0.99$,*
3. *$\mathbb{P}_{S \sim \mathcal{D}^m}[R(\mathit{ALG}(S)) \leq \delta] \geq 0.99$,*

then:

$$QC(\mathit{ALG}, T, m, \epsilon) \geq \log\left(\frac{\eta}{3 \cdot C \cdot \delta}\right).$$

The lower bound obtained in Theorem 4 is useful in situations when ALG(S) has high accuracy but the adversarial risk is large. This is a typical situation when using neural networks – one is often able to find classifiers that have high accuracy but they are not adversarially robust.

Summary: For a rich class of learning algorithms our security guarantee against query-bounded adversaries increases with accuracy. A risk of $2^{-\Omega(k)}$ leads to robustness against $\Theta(k)$-bounded adversaries.

8. Discussion

For a given task the QC can vary considerably depending on the learning algorithm. In this section we try to explain our current understanding of what governs this dependence.

[†] This lower bound is conditional on *Cap conjecture* (in the appendix).

[‡] This summary is conditional on *Cap conjecture*.

Consider the two intervals learning task from Theorem 2. As proven, if we use the 1-NN classifier the QC is lower-bounded by $\Theta(m)$, where m is the number of 'training' examples. Now consider what happens if we used the concept class of linear separators with the standard ERM algorithm. Then we expect the learned classifier to be a line that approximately separates the two intervals. To approximately recover this line the adversary can find two points through which the line passes by running two binary search procedures. This implies that the QC is independent of m. Thus QCs can be as different as $\Omega(m)$ and $O(1)$ depending on the used concept class/learning algorithm. Note that if we used the SVM classifier then with allowed perturbation of $z/10$ (as in Theorem 2) the adversarial risk will be 0 (for m big enough), as the learned classifier will be defined by the line passing through $(0, z/2)$ and $(m, z/2)$, which makes the QC degenerate and equal 0.

Consider now the CS data set and see what happens when you very the number of samples m and the learning algorithm. As we mentioned in Section 4, for CS K-NN might need as many as $m = 2^{\Theta(d \log d)}$ samples. This means that the lower bound of $\Theta(m)$ (from Theorem 2) becomes $2^{\Theta(d \log d)}$. On the other hand, QNN applied to this task is conjectured to have QC of $\Theta(d^2)$. We can also look at an interesting learning algorithm from Montasser et al. (2019), which for a concept class $\mathcal{H}$ can improperly learn $\mathcal{H}$ to a constant adversarial risk using $O(\text{VC-dim}(\mathcal{H}) \cdot \text{dualVC-dim}(\mathcal{H}))$ many samples. As the VC-dimension, and also the dual VC-dimension, of ellipsoids is in $O(d^2)$ we get that this algorithm achieves constant adversarial risk using $m = O(d^4)$ samples. The QC in this case is, in a way, irrelevant because the adversarial risk is small already. Thus we see three algorithms with different QCs but it's not clear if we can directly compare them as they require different sample sizes.

These examples suggest that there is a strong connection between the number of degrees of freedom of the classifier and the QC. However this connection cannot be expressed in terms of VC-dim or AdversarialVC-dim (Cullina et al., 2018) as there exist concept classes and learning tasks for which QC $\gg$ VC-dim/AdversarialVC-dim. Understanding the connection between the QC and the number of parameters is a part of ongoing work.

9. Conclusions and takeaways

We investigate robustness of learning algorithms against query-bounded adversaries. We start by introducing a definition of QC of adversarial attacks and then proceed to study it's properties. We prove a general lower bound for the QC in terms of an entropy-like property. Using this result we then show a series of lower bounds for classical learning algorithms. Specifically, we give a lower bound of $\Theta(d)$ for QNNs and a lower bound of $\Theta(m)$ for 1-NN algorithm. Moreover we observe that sources of the entropy can be varied. For K-NN the entropy is high due to the locality of the learning algorithm whereas for QNN it comes from the rotational symmetry of the data. The entropy can also be increased by introducing randomness in the learning algorithm itself. We also show that improvements in accuracy of a model lead to an improved security against query-bounded adversaries.

Our analysis identifies properties of learning algorithms that make them (non-)robust. These results give a rule-of-thumb: "The higher the entropy of decision boundary the better" for assessing the QC of a given algorithm.

We believe that a systematic investigation of learning algorithms from the point of view of QC will lead to more adversarially-robust systems. Specifically, it should be possible to design generic defenses that can be applied on top of any learning algorithm. One example of such a defense was given in Section 6. Significantly more work is needed in order to fulfill the potential of this approach. But imagine that this type of defense could be applied efficiently with only a black-box access to the underlying classifier. And imagine further, that it could guaranteed a QC of, say $q = 2^{\Theta(d)}$. This would arguably solve the adversarial robustness problem.

References

Ashtiani, H., Pathak, V., and Urner, R. Black-box certification and learning under adversarial perturbations. *ArXiv*, abs/2006.16520, 2020.

Athalye, A., Carlini, N., and Wagner, D. A. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 274–283, 2018. URL http://proceedings.mlr.press/v80/athalye18a.html.

Bhagoji, A., Cullina, D., and Mittal, P. Dimensionality reduction as a defense against evasion attacks on machine learning classifiers. 04 2017.

Bhambri, S., Muku, S., Tulasi, A. S., and Buduru, A. B. A study of black box adversarial attacks in computer vision. *ArXiv*, abs/1912.01667, 2019.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In Blockeel, H., Kersting, K., Nijssen, S., and Železný, F. (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp.

387–402, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40994-3.

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. P. Adversarial examples from cryptographic pseudo-random generators. *CoRR*, abs/1811.06418, 2018. URL http://arxiv.org/abs/1811.06418.

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. Adversarial examples from computational constraints. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 831–840, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/bubeck19a.html.

Carlini, N. and Wagner, D. A. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, pp. 3–14, 2017. doi: 10.1145/3128572.3140444. URL https://doi.org/10.1145/3128572.3140444.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D. Adversarial attacks and defences: A survey. *ArXiv*, abs/1810.00069, 2018.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, pp. 15–26, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450352024. doi: 10.1145/3128572.3140448. URL https://doi.org/10.1145/3128572.3140448.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/cohen19c.html.

Collobert, R. and Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 160–167, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390177. URL http://doi.acm.org/10.1145/1390156.1390177.

Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, 13:21–27, 1967.

Cullina, D., Bhagoji, A. N., and Mittal, P. Pac-learning in the presence of evasion adversaries. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 228–239, Red Hook, NY, USA, 2018. Curran Associates Inc.

Diochnos, D. I., Mahloujifar, S., and Mahmoody, M. Adversarial risk and robustness: General definitions and implications for the uniform distribution. *ArXiv*, abs/1810.12272, 2018.

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., and Goodfellow, I. J. Adversarial spheres. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*, 2018. URL https://openreview.net/forum?id=SkthlLkPf.

Gluch, G. and Urbanke, R. L. Constructing a provably adversarially-robust classifier from a high accuracy one. *CoRR*, abs/1912.07561, 2019. URL http://arxiv.org/abs/1912.07561.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. 2014. URL http://arxiv.org/abs/1412.6572. cite arxiv:1412.6572.

Gourdeau, P., Kanade, V., Kwiatkowska, M., and Worrell, J. On the hardness of robust classification. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/8133415ea4647b6345849fb38311cf32-Paper.pdf.

Hayes, J. and Danezis, G. Machine learning as an adversarial service: Learning black-box adversarial examples. 08 2017.

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. Black-box adversarial attacks with limited queries and information. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2137–2146, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/ilyas18a.html.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In

Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012.

Lécuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pp. 656–672, 2019. doi: 10.1109/SP.2019.00044. URL https://doi.org/10.1109/SP.2019.00044.

Liu, Y., Chen, X., Liu, C., and Song, D. Delving into transferable adversarial examples and black-box attacks. *CoRR*, abs/1611.02770, 2016. URL http://arxiv.org/abs/1611.02770.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Montasser, O., Hanneke, S., and Srebro, N. Vc classes are adversarially robustly learnable, but only improperly. volume 99 of *Proceedings of Machine Learning Research*, pp. 2512–2530, Phoenix, USA, 25–28 Jun 2019. PMLR. URL http://proceedings.mlr.press/v99/montasser19a.html.

Nguyen, A. M., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pp. 427–436. IEEE Computer Society, 2015. ISBN 978-1-4673-6964-0. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#NguyenYC15.

Papernot, N., McDaniel, P., and Goodfellow, I. J. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *ArXiv*, abs/1605.07277, 2016.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. pp. 506–519, 04 2017. doi: 10.1145/3052973.3053009.

Salman, H., Yang, G., Li, J., Zhang, P., Zhang, H., Razenshteyn, I. P., and Bubeck, S. Provably robust deep learning via adversarially trained smoothed classifiers. *ArXiv*, abs/1906.04584, 2019.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 5019–5031, Red Hook, NY, USA, 2018. Curran Associates Inc.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6199.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkZvSe-RZ.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

Uesato, J., O'Donoghue, B., Kohli, P., and van den Oord, A. Adversarial risk and the dangers of evaluating against weak attacks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 5032–5041, 2018. URL http://proceedings.mlr.press/v80/uesato18a.html.

Wang, Y., Jha, S., and Chaudhuri, K. Analyzing the robustness of nearest neighbors to adversarial examples. volume 80 of *Proceedings of Machine Learning Research*, pp. 5133–5142, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/wang18c.html.

Xiao, C., Li, B., Zhu, J.-Y., He, W., Liu, M., and Song, D. Generating adversarial examples with adversarial networks, 2018. URL https://openreview.net/forum?id=HknbyQbC-.

Xiao, K. Y., Tjeng, V., Shafiullah, N. M. M., and Madry, A. Training for faster adversarial robustness verification via inducing reLU stability. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJfIVjAcKm.

Xu, W., Evans, D., and Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. 04 2017.

Yin, D., Kannan, R., and Bartlett, P. Rademacher complexity for adversarially robust generalization. volume 97 of *Proceedings of Machine Learning Research*, pp. 7085–7094, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/yin19b.html.

Zheng, T., Chen, C., and Ren, K. Distributionally adversarial attack. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:2253–2260, 07 2019. doi: 10.1609/aaai.v33i01.33012253.