Analysis of classifiers' robustness to adversarial perturbations

Alhussein Fawzi* Omar Fawzi† Pascal Frossard*

Abstract

The robustness of a classifier to arbitrary small perturbations of the datapoints is a highly desirable property when the classifier is deployed in real and possibly hostile environments. In this paper, we propose a theoretical framework for analyzing the robustness of classifiers to adversarial perturbations, and study two common families of classifiers. In both cases, we show the existence of a fundamental limit on the robustness to adversarial perturbations, which is expressed in terms of a *distinguishability* measure between the classes. Our result implies that in tasks involving small distinguishability, *no classifier* will be robust to adversarial perturbations, even if a good accuracy is achieved. Furthermore, we show that robustness to random noise does not imply, in general, robustness to adversarial perturbations. In fact, in high dimensional problems, linear classifiers are shown to be much more robust to random noise than to adversarial perturbations. Our analysis is complemented by experimental results on controlled and real-world data. Up to our knowledge, this is the first theoretical work that addresses the surprising phenomenon of adversarial instability recently observed for deep networks Szegedy et al. (2014). Our work shows that this phenomenon is not limited to deep networks, and gives a theoretical explanation of the causes underlying the adversarial instability of classifiers.

1 Introduction

In most classification settings, the proportion of misclassified samples in the test set is the main performance metric used to evaluate classifiers. However, the robustness of classifiers to small arbitrary perturbations of datapoints is also a highly desirable property, when the classifier is deployed in real, and possibly hostile environments. In particular, we expect that any slight perturbation of the datapoints does not change dramatically the classifier's prediction. The goal of this paper is to study the robustness of classifiers to *adversarial perturbations*: given a classifier f and a datapoint x, what is the norm of the *minimal* perturbation r such that $x + r$ is classified differently than x? This quantity, averaged over all datapoints, is called *the robustness of f to adversarial perturbations*. The lack of robustness to small adversarial perturbations can be problematic, as an attacker having the knowledge of the classifier's parameters will only need to apply a slight (often imperceptible) perturbation to the data to change the classifier's decision. In addition, the robustness to adversarial perturbations of f can also be used to assess the quality of the features that are used to distinguish between the classes. For instance, given an "airplane" vs. "car" image classification task, we expect that the smallest perturbation r needed to transform the image of a car to one of a plane to be large. Therefore, if we seek to have a classifier that behaves according to our common understanding of airplanes and cars, it should have a large robustness to adversarial perturbations.

In this paper, we introduce a framework for formally studying the robustness of classifiers to adversarial perturbations in the binary setting. The robustness properties of linear and quadratic classifiers are studied in detail. In both cases, our results show the existence of a fundamental limit on the robustness to adversarial perturbations. This limit is expressed in terms of a *distinguishability* measure between the classes, which depends on the considered family of classifiers. Specifically, for linear classifiers, the distinguishability is defined as the distance between the means of the two classes, while for quadratic classifiers, it is defined as the distance between the matrices of second order moments of the two classes. Our upper bound on the robustness is valid for all classifiers independently of the training procedure. This result has the following important implication: in practical classification tasks involving a small value of distinguishability, any linear or quadratic classifier with low misclassification rate *will not be robust to adversarial perturbations*. We further compare the robustness to adversarial perturbations of linear classifiers to the more traditional notion of robustness to random uniform noise. In high dimensions, the latter robustness is shown to be much larger than the former, thereby showing a fundamental difference between the two notions of robustness. In fact, in high dimensional classification tasks, linear classifiers can be robust to random noise, even for small values of the distinguishability. We illustrate the newly introduced concepts and our theoretical results on a running example used throughout the paper, as well as practical classification tasks.

Our paper is motivated by an important recent contribution (Szegedy et al., 2014) that shows empirically that state-of-the-art deep networks are surprisingly unstable to hardly perceptible adversarial perturbations. It was suggested that the instability of these classifiers is due to the high *nonlinearity* of neural networks, but no theoretical justification was provided. The work in Szegedy et al. (2014) received a widespread interest from the press, as the instability of deep

*Ecole Polytechnique Federale de Lausanne (EPFL), Signal Processing Laboratory (LTS4), Lausanne 1015-Switzerland. Email: (alhussein.fawzi@epfl.ch, pascal.frossard@epfl.ch)

†ENS de Lyon, LIP, UMR 5668 ENS Lyon - CNRS - UCBL - INRIA, Université de Lyon, France. Email: omar.fawzi@ens-lyon.fr

networks to small adversarial perturbations raises significant challenges and their cause was a mystery. In fact, how is it possible that state-of-the-art deep networks generalize well on the test set, but are unable to classify correctly points that are hardly distinguishable from training data? Our paper shows that there can be cases where the classifier is perfectly accurate, but arbitrarily small perturbations can flip the decision of the classifier. Moreover, we show theoretically that adversarial instability is not limited to neural networks, but is a much more general phenomenon.

Following the original paper of Szegedy et al. (2014), several attempts have been made to make deep networks robust to adversarial perturbations Chalupka et al. (2014); Gu & Rigazio (2014), and a related phenomenon has been explored in Nguyen et al. (2014). In a very recent and independent work, Goodfellow et al. (2014) argue that instability of neural networks is mainly due to their *linear* nature in high dimensions. While this hypothesis is contrary to previous explanations that invoked the highly *nonlinear* nature of neural networks Szegedy et al. (2014), our theoretical results go in the same direction and suggest a more general trend: in difficult classification problems involving small measures of distinguishability, a classifier that is not "flexible" enough for the task will not be robust to adversarial perturbations (even if a low risk is achieved).

Our work should not be confused with works on the security of machine learning algorithms under adversarial attacks (Biggio et al., 2012; Barreno et al., 2006; Dalvi et al., 2004). These works specifically study attacks that manipulate *the learning system* (e.g., change the decision function by injecting malicious training points), as well as defense strategies to counter these attacks. This setting significantly differs from ours, as we examine the robustness of *a fixed* classifier to adversarial perturbations (that is, the classifier cannot be manipulated). Finally, the stability of learning algorithms has been defined and extensively studied in (Bousquet & Elisseeff, 2002; Lugosi & Pawlak, 1994). Again, this notion of stability differs from the one studied here, as we are interested in the robustness of fixed classifiers, and not of learning algorithms.

The paper is structured as follows: Sec. 2 introduces the problem setting. In Sec. 3, we introduce a running example that is used throughout the paper. The robustness of linear classifiers (to adversarial and random noise) is studied in Sec. 4. In Sec. 5, we study quadratic classifiers. Sec. 6 presents experiments on real data; conclusions and outlook on future reseach are given in Sec. 7. Besides, all the proofs can be found in the appendix.

2 Problem setting

We first introduce the framework and notations that are used for analyzing the robustness of classifiers to adversarial and uniform random noise. We restrict our analysis to the binary classification task, for simplicity. We expect similar conclusions for the multi-class case, but we leave that for future work. We let μ denote the probability measure on $\mathbb{R}^d$ of the data points we wish to classify, and $y(x) \in \{-1, 1\}$ be the label of a point $x \in \mathbb{R}^d$. The distribution μ is assumed to be of bounded support. That is, $\mathbb{P}_\mu(\|x\|_2 \leq M) = 1$, for some $M > 0$. We denote by μ_1 and μ_{-1} the distributions of class 1 and class -1 in $\mathbb{R}^d$, respectively. Let $f : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary classification function. The classification rule associated to f is simply obtained by taking the sign of $f(x)$. The performance of a classifier f is usually measured through its *risk*, defined by the probability of misclassification according to μ:

$$\begin{aligned} R(f) &= \mathbb{P}_\mu(\text{sign}(f(x)) \neq y(x)) \\ &= p_1 \mathbb{P}_{\mu_1}(f(x) < 0) + p_{-1} \mathbb{P}_{\mu_{-1}}(f(x) \geq 0), \end{aligned}$$

where $p_{\pm 1} = \mathbb{P}_\mu(y(x) = \pm 1)$.

The focus of this paper is to study the robustness of classifiers to adversarial perturbations in the ambient space $\mathbb{R}^d$. Given a datapoint $x \in \mathbb{R}^d$ sampled from μ, we denote by $\Delta_{\text{adv}}(x; f)$ the norm of the smallest perturbation that switches the sign[1] of f:

$$\Delta_{\text{adv}}(x; f) = \min_{r \in \mathbb{R}^d} \|r\|_2 \text{ subject to } f(x) f(x + r) \leq 0. \tag{1}$$

Unlike random noise, the above definition corresponds to a minimal noise, where the perturbation r is sought to flip the estimated label of x. This justifies the *adversarial* nature of the perturbation. It is important to note that, while x is a datapoint sampled according to μ, the perturbed point $x + r$ is not required to belong to the dataset (i.e., $x + r$ can be outside the support of μ). The robustness to adversarial perturbation of f is defined as the average of $\Delta_{\text{adv}}(x; f)$ over all x:

$$\rho_{\text{adv}}(f) = \mathbb{E}_\mu(\Delta_{\text{adv}}(x; f)). \tag{2}$$

In words, $\rho_{\text{adv}}(f)$ is defined as the average norm of the minimal perturbations required to flip the estimated labels of the datapoints. Note that $\rho_{\text{adv}}(f)$ is a property of the classifier f and the distribution μ, but is independent of the true labels of the datapoints y.[2]

[1] We make the assumption that a perturbation r that satisfies the equality $f(x + r) = 0$ flips the estimated label of x.

[2] In that aspect, our definition slightly differs from the one proposed in Szegedy et al. (2014), which defines the robustness to adversarial perturbations as the average norm of the minimal perturbation required to *misclassify* all datapoints. Our notion of robustness is larger than theirs; our upper bounds therefore also directly apply for their definition of robustness.

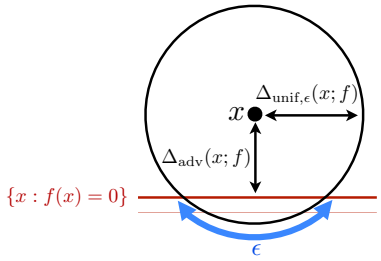


Figure 1: Illustration of $\Delta_{\text{adv}}(x; f)$ and $\Delta_{\text{unif},\epsilon}(x; f)$. The red line represents the classifier boundary. In this case, the quantity $\Delta_{\text{adv}}(x; f)$ is equal to the distance from x to this line. The radius of the sphere drawn around x is $\Delta_{\text{unif},\epsilon}(x; f)$. Assuming $f(x) > 0$, observe that the spherical cap in the region below the line has measure ϵ, which means that the probability that a random point sampled on the sphere has label $+1$ is $1 - \epsilon$.

Quantity	Dependence
$R(f) = \mathbb{P}_\mu(\text{sign}(f(x)) \neq y(x))$	μ, y, f
$\rho_{\text{adv}}(f) = \mathbb{E}_\mu(\Delta_{\text{adv}}(x; f))$	μ, f
$\rho_{\text{unif},\epsilon}(f) = \mathbb{E}_\mu(\Delta_{\text{unif},\epsilon}(x; f))$	μ, f

Table 1: Quantities of interest: risk, robustness to adversarial perturbations, and robustness to random uniform noise, respectively.

In this paper, we also study the robustness of classifiers to random uniform noise, that we define as follows. For a given $\epsilon \in [0, 1]$, let

$$\Delta_{\text{unif},\epsilon}(x; f) = \max_{\eta \geq 0} \eta \quad \text{s.t. } \mathbb{P}_{n \sim \eta\mathbb{S}}(f(x)f(x+n) \leq 0) \leq \epsilon, \tag{3}$$

where $\eta\mathbb{S}$ denotes the uniform measure on the sphere centered at 0 and of radius η in $\mathbb{R}^d$. In words, $\Delta_{\text{unif},\epsilon}(x; f)$ denotes the maximal radius of the sphere centered at x, such that perturbed points sampled uniformly at random from this sphere are classified similarly to x with high probability. An illustration of $\Delta_{\text{unif},\epsilon}(x; f)$ and $\Delta_{\text{adv}}(x; f)$ is given in Fig. 1. Similarly to adversarial perturbations, the point $x + n$ will lie outside the support of μ, in general. Note moreover that $\Delta_{\text{unif},\epsilon}(x; f)$ provides an upper bound on $\Delta_{\text{adv}}(x; f)$, for all ϵ. The ϵ-robustness of f to random uniform noise is defined by:

$$\rho_{\text{unif},\epsilon}(f) = \mathbb{E}_\mu(\Delta_{\text{unif},\epsilon}(x; f)). \tag{4}$$

We summarize the quantities of interest in Table 1.

3 Running example

We introduce in this section a running example used throughout the paper to illustrate the notion of adversarial robustness, and highlight its difference with the notion of risk. We consider a binary classification task on square images of size $\sqrt{d} \times \sqrt{d}$. Images of class 1 (resp. class -1) contain exactly one vertical line (resp. horizontal line), and a small constant positive number a (resp. negative number $-a$) is added to all the pixels of the images. That is, for class 1 (resp. -1) images, background pixels are set to a (resp. $-a$), and pixels belonging to the line are equal to $1 + a$ (resp. $1 - a$). Fig. 2 illustrates the classification problem for $d = 25$. The number of datapoints to classify is $N = 2\sqrt{d}$. Clearly, the most visual concept that permits to separate the two classes is the *orientation* of the line (i.e., horizontal vs. vertical). The *bias* of the image (i.e., the sum of all its pixels) is also a valid concept for this task, as it separates the two classes, despite being much more difficult to detect visually. The class of an image can therefore be correctly estimated from its orientation *or* from the bias. The linear classifier defined by

$$f_{\text{lin}}(x) = \frac{1}{\sqrt{d}} \mathbf{1}^T x - 1, \tag{5}$$

where $\mathbf{1}$ is the vector of size d whose entries are all equal to 1, and x is the vectorized image, exploits the difference of bias between the two classes and achieves a perfect classification accuracy for all $a > 0$. Indeed, a simple computation gives $f_{\text{lin}}(x) = \sqrt{d}a$ (resp. $f_{\text{lin}}(x) = -\sqrt{d}a$) for class 1 (resp. class -1) images. Therefore, the risk of f_{lin} is $R(f_{\text{lin}}) = 0$. It is important to note that f_{lin} only achieves zero risk because it captures the bias, but fails to distinguish between the images from the orientation of the line. Indeed, when $a = 0$, the datapoints are not linearly separable. Despite its perfect accuracy for any $a > 0$, f_{lin} is *not* robust to small adversarial perturbations when a is small, as a minor perturbation of

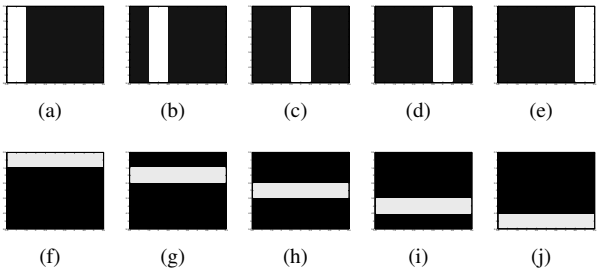


Figure 2: (a...e): Class 1 images. (f...j): Class -1 images.

Figure 3: Robustness to adversarial noise of linear and quadratic classifiers. (a): Original image, (b,c): Minimally perturbed image that switches the estimated label of (b) f_{lin}, (c) f_{quad}. Note that the difference between (b) and (a) is hardly perceptible, this demonstrates that f_{lin} is not robust to adversarial noise. On the other hand images (c) and (a) are clearly different, which indicates that f_{quad} is more robust to adversarial noise. Parameters: $d = 4$, and $a = 0.1/\sqrt{d}$.

the bias switches the estimated label. Indeed, a simple computation gives $\rho_{\text{adv}}(f_{\text{lin}}) = \sqrt{d}a$; therefore, the adversarial robustness of f_{lin} can be made arbitrarily small by choosing a small enough a. More than that, among all linear classifiers that satisfy $R(f) = 0$, f_{lin} is the one that maximizes $\rho_{\text{adv}}(f)$ (as we show later in Section 4). Therefore, *all* zero-risk linear classifiers are not robust to adversarial perturbations, for this task. Unlike linear classifiers, a more *flexible* classifier that correctly captures the orientation will be robust to adversarial perturbation, unless this perturbation significantly alters the image and modifies the direction of the line. To illustrate this point, we compare the adversarial robustness of f_{lin} to that of a second order polynomial classifier f_{quad} that achieves zero risk in Fig. 3, for $d = 4$.[3] While a hardly perceptible change of the image is enough to switch the estimated label for the linear classifier, the minimal perturbation for f_{quad} is one that modifies the direction of the line, to a great extent.

The above example highlights several important facts, that we summarize as follows:

- **Risk and adversarial robustness are two distinct properties of a classifier.** While $R(f_{\text{lin}}) = 0$, f_{lin} is definitely not robust to small adversarial perturbations.[4] This is due to the fact that f_{lin} only captures the bias, and ignores the orientation of the line.
- **To capture orientation (i.e., the most visual concept), one has to use a classifier that is flexible enough for the task.** Unlike the class of linear classifiers, the class of polynomial classifiers of degree 2 correctly captures the line orientation, for $d = 4$.
- **The robustness to adversarial perturbations provides a quantitative measure of the strength of a concept.** Since $\rho_{\text{adv}}(f_{\text{lin}}) \ll \rho_{\text{adv}}(f_{\text{quad}})$, one can confidently say that the concept captured by f_{quad} is *stronger* than that of f_{lin}, in the sense that the essence of the classification task is captured by f_{quad}, but not by f_{lin} (while they are equal in terms of misclassification rate). In general classification problems, the quantity $\rho_{\text{adv}}(f)$ provides a natural way to evaluate and compare the learned concept; larger values of $\rho_{\text{adv}}(f)$ indicate that stronger concepts are learned, for comparable values of the risk.

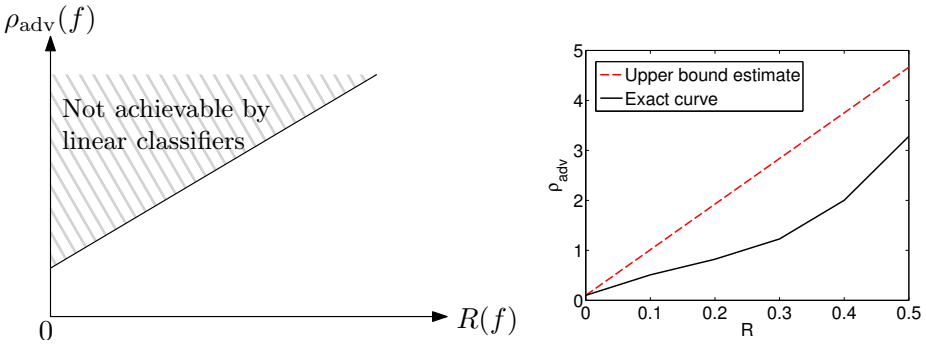
Similarly to the above example, we believe that the robustness to adversarial perturbations is key to assess the strength of a concept, in real-world classification tasks. In these cases, weak concepts will correspond to partial information about the classification task (which are possibly enough to achieve a good accuracy), while strong concepts will capture the essence of the classification task. We study in the next sections the robustness of two classes of classifiers to adversarial perturbations.

4 Linear classifiers

We study in this section the robustness of linear classifiers to adversarial perturbations, and uniform random noise.

[3] We postpone the detailed analysis of f_{quad} to Section 5.

[4] The opposite is also possible, since a constant classifier (e.g., $f(x) = 1$ for all x) is clearly robust to perturbations, but does not achieve good accuracy.



(a) (b)

Figure 4: ρ_{adv} versus risk diagram for linear classifiers. Each point in the plane represents a linear classifier f. (a): Illustrative diagram, with the non-achievable zone (Theorem 4.1). (b): The exact ρ_{adv} versus risk achievable curve, and our upper bound estimate on the running example.

4.1 Adversarial perturbations

We define the classification function $f(x) = w^T x + b$. In this case, the adversarial perturbation function $\Delta_{\text{adv}}(x; f)$ can be computed in closed form and is equal to the distance from x to the hyperplane $\{f(x) = 0\}$: $\Delta_{\text{adv}}(x; f) = |w^T x + b|/\|w\|_2$. Note that any linear classifier for which $|b| > M\|w\|_2$ is a trivial classifier that assigns the same label to all points, and we therefore assume that $|b| \leq M\|w\|_2$. The following theorem bounds $\rho_{\text{adv}}(f)$ from above in terms of the first moments of the distributions μ_1 and μ_{-1}, and the classifier's risk:

Theorem 4.1. *Let $f(x) = w^T x + b$ such that $|b| \leq M\|w\|_2$. Then,*

$$\rho_{adv}(f) \leq \|p_1 \mathbb{E}_{\mu_1}(x) - p_{-1}\mathbb{E}_{\mu_{-1}}(x)\|_2 + M(|p_1 - p_{-1}| + 4R(f)).$$

In the balanced setting where $p_1 = p_{-1} = 1/2$, and if the intercept $b = 0$ the following inequality holds:

$$\rho_{adv}(f) \leq \frac{1}{2}\|\mathbb{E}_{\mu_1}(x) - \mathbb{E}_{\mu_{-1}}(x)\|_2 + 2MR(f).$$

Our upper bound on $\rho_{\text{adv}}(f)$ depends on the difference of means $\|\mathbb{E}_{\mu_1}(x) - \mathbb{E}_{\mu_{-1}}(x)\|_2$, which measures the distinguishability between the classes. Note that this term is classifier-independent, and is only a property of the classification task. The only dependence on f in the upper bound is through the risk $R(f)$. Thus, in classification tasks where the means of the two distributions are close (i.e., $\|\mathbb{E}_{\mu_1}(x) - \mathbb{E}_{\mu_{-1}}(x)\|_2$ is small), *any linear classifier* with small risk will necessarily have a small robustness to adversarial perturbations. Note that the upper bound logically increases with the risk, as there clearly exist robust linear classifiers that achieve high risk (e.g., constant classifier). Fig. 4 (a) pictorially represents the ρ_{adv} vs R diagram as predicted by Theorem 4.1. Each linear classifier is represented by a point on the ρ_{adv}–R diagram, and our result shows the existence of a region that linear classifiers cannot attain.

Quite importantly, in many interesting classification problems, the quantity $\|\mathbb{E}_{\mu_1}(x) - \mathbb{E}_{\mu_{-1}}(x)\|_2$ is small due to large intra-class variability (e.g., due to complex intra-class geometric transformations in computer vision applications). Therefore, even if a linear classifier can achieve a good classification performance on such a task, it will not be robust to small adversarial perturbations. In simple tasks involving distributions with significantly different averages, it is likely that there exists a linear classifier that can separate correctly the classes, and have a large robustness to adversarial perturbations.

4.2 Random uniform noise

We now examine the robustness of linear classifiers to random uniform noise. The following theorem compares the robustness of linear classifiers to random uniform noise, with the robustness to adversarial perturbations.

Theorem 4.2. *Let $f(x) = w^T x + b$. For any $\epsilon \in [0, 1/12)$, we have the following bounds on $\rho_{unif,\epsilon}(f)$:*

$$\rho_{unif,\epsilon}(f) \geq \max\left(C_1(\epsilon)\sqrt{d}, 1\right)\rho_{adv}(f), \tag{6}$$

$$\rho_{unif,\epsilon}(f) \leq \widetilde{C_2}(\epsilon, d)\rho_{adv}(f) \leq C_2(\epsilon)\sqrt{d}\rho_{adv}(f), \tag{7}$$

with $C_1(\epsilon) = (2\ln(2/\epsilon))^{-1/2}$, $\widetilde{C_2}(\epsilon, d) = (1 - (12\epsilon)^{1/d})^{-1/2}$ and $C_2(\epsilon) = (1 - 12\epsilon)^{-1/2}$.

In words, $\rho_{\text{unif},\epsilon}(f)$ behaves as $\sqrt{d}\rho_{\text{adv}}(f)$ for linear classifiers (for constant ϵ). Linear classifiers are therefore more robust to random noise than adversarial perturbations, by a factor of $\sqrt{d}$. In typical high dimensional classification

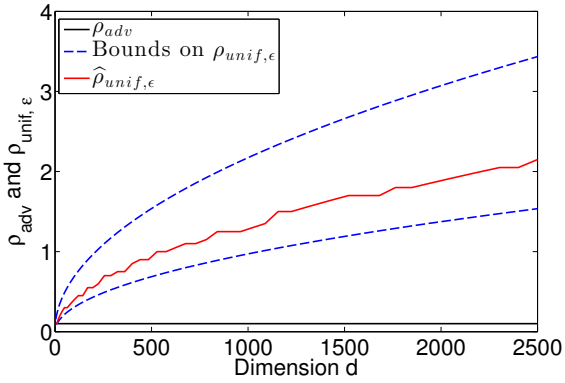


Figure 5: Adversarial robustness and robustness to random uniform noise of f_{lin} versus the dimension d. We used $\epsilon = 0.01$, and $a = 0.1/\sqrt{d}$. The lower bound is given in Eq. (6), and the upper bound is the first inequality in Eq. (7).

problems, this shows that a linear classifier can be robust to random noise even if $\|\mathbb{E}_{\mu_1}(x) - \mathbb{E}_{\mu_{-1}}(x)\|_2$ is small. Note moreover that our result is tight for $\epsilon = 0$, as we get $\rho_{\text{unif},0}(f) = \rho_{\text{adv}}(f)$.

Our results can be put in perspective with the empirical results of Szegedy et al. (2014), that showed a large gap between the two notions of robustness on neural networks. Our analysis provides a confirmation of this high dimensional phenomenon on linear classifiers.

4.3 Example

We now illustrate our theoretical results on the example of Section 3. In this case, we have $\|\mathbb{E}_{\mu_1}(x) - \mathbb{E}_{\mu_{-1}}(x)\|_2 = 2\sqrt{d}a$. By using Theorem 4.1, *any* zero-risk linear classifier satisfies $\rho_{\text{adv}}(f) \leq \sqrt{d}a$. As we choose $a \ll 1/\sqrt{d}$, accurate linear classifiers are therefore not robust to adversarial perturbations, for this task. We note that f_{lin} (defined in Eq.(5)) achieves the upper bound and is therefore the more robust accurate linear classifier one can get, as it can easily be checked that $\rho_{\text{adv}}(f_{\text{lin}}) = \sqrt{d}a$. In Fig. 4 (b) the exact ρ_{adv} vs R curve is compared to our theoretical upper bound[5], for $d = 25$, $N = 10$ and a bias $a = 0.1/\sqrt{d}$. Besides the zero-risk case where our upper bound is tight, the upper bound is reasonably close to the exact curve for other values of the risk (despite not being tight).

We now focus on the robustness to uniform random noise of f_{lin}. For various values of d, we compute the upper and lower bounds on the robustness to random uniform noise (Theorem 4.2) of f_{lin}, where we fix ϵ to 0.01. In addition, we compute a simple empirical estimate $\widehat{\rho}_{\text{unif},\epsilon}$ of the robustness to random uniform noise of f_{lin} (see Sec. 6 for details on the computation of this estimate). The results are illustrated in Fig. 5. While the adversarial noise robustness is constant with the dimension (equal to 0.1, as $\rho_{\text{adv}}(f_{\text{lin}}) = \sqrt{d}a$ and $a = 0.1/\sqrt{d}$), the robustness to random uniform noise *increases* with d. For example, for $d = 2500$, the value of $\rho_{\text{unif},\epsilon}$ is at least 15 times larger than adversarial robustness ρ_{adv}. In high dimensions, a linear classifier is therefore much more robust to random uniform noise than adversarial noise.

5 Quadratic classifiers

5.1 Analysis of adversarial perturbations

We study the robustness to adversarial perturbations of quadratic classifiers of the form $f(x) = x^T Ax$, where A is a symmetric matrix. Besides the practical use of quadratic classifiers in some applications (Goldberg & Elhadad, 2008; Chang et al., 2010), they represent a natural extension of linear classifiers. The study of linear vs. quadratic classifiers provides insights into how adversarial robustness depends on the family of considered classifiers. Similarly to the linear setting, we exclude the case where f is a trivial classifier that assigns a constant label to all datapoints. That is, we assume that A satisfies

$$\lambda_{\min}(A) < 0, \quad \lambda_{\max}(A) > 0, \tag{8}$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the smallest and largest eigenvalues of A. We moreover impose that the eigenvalues of A satisfy

$$\max\left(\left|\frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}\right|, \left|\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}\right|\right) \leq K, \tag{9}$$

for some constant value $K \geq 1$ (independent of matrix A). This assumption imposes an approximate symmetry around 0 of the extremal eigenvalues of A, thereby disallowing a large bias towards any of the two classes. The following result

[5]The exact curve is computed using a bruteforce approach (we omit the details for space constraints).

bounds the adversarial robustness of quadratic classifiers as a function of the second order moments of the distribution and the risk.

Theorem 5.1. *Let $f(x) = x^T Ax$, where A satisfies Eqs.* (8) *and* (9)*. Then,*

$$\rho_{adv}(f) \leq 2\sqrt{K\|p_1 C_1 - p_{-1} C_{-1}\|_* + 2MKR(f)},$$

where $C_{\pm 1}(i,j) = (\mathbb{E}_{\mu_{\pm 1}}(x_i x_j))_{1 \leq i,j \leq d}$, and $\|\cdot\|_*$ denotes the nuclear norm defined as the sum of the singular values of the matrix.

In words, the upper bound on the adversarial robustness depends on a distinguishability measure, defined by $\|C_1 - C_{-1}\|_*$, and the classifier's risk. In difficult classification tasks, where $\|C_1 - C_{-1}\|_*$ is small, any quadratic classifier with low risk and satisfying our assumptions is not robust to adversarial perturbations.

It should be noted that, while the distinguishability was measured with the distance between the means of the two distributions in the linear case, it is defined here as the difference between the second order moments matrices $\|C_1 - C_{-1}\|_*$. Therefore, in classification tasks involving two distributions with close means, and different second order moments, any zero-risk linear classifier will not be robust to adversarial noise, while zero-risk and robust quadratic classifiers are a priori possible according to our upper bound in Theorem 5.1. This suggests that robustness to adversarial perturbations can be larger for more flexible classifiers, for comparable values of the risk.

5.2 Example

We now illustrate our results on the running example of Section 3, with $d = 4$. In this case, a simple computation gives $\|C_1 - C_{-1}\|_* = 2 + 8a \geq 2$. This term is significantly larger than the difference of means (equal to $4a$), and there is therefore hope to have a quadratic classifier that is accurate *and* robust to small adversarial perturbations, according to Theorem 5.1. In fact, the following quadratic classifier

$$f_{\text{quad}}(x) = x_1 x_2 + x_3 x_4 - x_1 x_3 - x_2 x_4,$$

outputs 1 for vertical images, and -1 for horizontal images (independently of the bias a). Therefore, f_{quad} achieves zero risk on this classification task, similarly to f_{lin}. The two classifiers however have different robustness properties to adversarial perturbations. Using straightforward calculations, it can be shown that $\rho_{\text{adv}}(f_{\text{quad}}) = 1/\sqrt{2}$, for any value of a (see Appendix D for more details). For small values of a, we therefore get $\rho_{\text{adv}}(f_{\text{lin}}) \ll \rho_{\text{adv}}(f_{\text{quad}})$. This result is intuitive, as f_{quad} differentiates the images from their *orientation*, unlike f_{lin} that uses the *bias* to distinguish them. The minimal perturbation required to switch the estimated label of f_{quad} is therefore one that modifies the direction of the line, while a hardly perceptible perturbation that modifies the bias is enough to flip the label for f_{quad}. Fig. 3 in Section 3 illustrates this result.

6 Experimental results

In this section, we illustrate our results on practical classification examples. Specifically, through experiments on real data, we seek to confirm the identified limit on the robustness of linear and quadratic classifiers, and we show the large gap between adversarial and random robustness on real data. We also study more general classifiers to suggest that the trends obtained with our theoretical results are not limited to linear and quadratic classifiers.

We perform experiments on several classifiers: linear SVM (denoted *L-SVM*), SVM with polynomial kernels of degree q (denoted *poly-SVM (q)*), and SVM with RBF kernel with a width parameter σ^2 (*RBF-SVM(σ^2)*). To train the classifiers, we use the efficient Liblinear Fan et al. (2008) and LibSVM Chang & Lin (2011) implementations, and we fix the regularization parameters using a cross-validation procedure. Given a classifier f, and a datapoint x, we use an approach close to that of Szegedy et al. (2014) to approximate $\Delta_{\text{adv}}(x; f)$. Specifically, we perform a line search to find the maximum $c > 0$ for which the minimizer of the following problem satisfies $f(x)f(x+r) \leq 0$:

$$\min_r c\|r\|_2 + L(f(x+r)\text{sign}(f(x))),$$

where we set $L(x) = \max(0, x)$. The above problem (for c fixed) is solved with a subgradient descent procedure, and we denote by $\widehat{\Delta}_{\text{adv}}(x; f)$ the obtained solution.[6] The empirical robustness to adversarial perturbations is then defined by $\widehat{\rho}_{\text{adv}}(f) = \frac{1}{m}\sum_{i=1}^m \widehat{\Delta}_{\text{adv}}(x_i; f)$, where $x_1, \ldots, x_m$ denote the training points. To evaluate the robustness of f, we compare $\widehat{\rho}_{\text{adv}}(f)$ to the following quantity:

$$\kappa = \frac{1}{m}\sum_{i=1}^m \min_{j: y(x_j) \neq y(x_i)} \|x_i - x_j\|_2.$$

[6] This procedure is not guaranteed to provide the optimal solution (for arbitrary classifiers f), as the problem is clearly non convex. Strictly speaking, the optimization procedure is only guaranteed to provide an upper bound on $\Delta_{\text{adv}}(x; f)$.

Model	Train error (%)	Test error (%)	$\widehat{\rho}_{\text{adv}}$	$\widehat{\rho}_{\text{unif},\epsilon}$
L-SVM	4.8	7.0	0.08	0.97
poly-SVM(2)	0	1	0.19	2.15
poly-SVM(3)	0	0.6	0.24	2.51
RBF-SVM(1)	0	1.1	0.16	-
RBF-SVM(0.1)	0	0.5	0.32	-

Table 2: Training and testing accuracy of different models, and robustness to adversarial noise for the MNIST task. Note that for this example, we have $\kappa = 0.72$.

It represents the average norm of the minimal perturbation required to "transform" a training point to a training point of the opposite class, and can be seen as a distance measure between the two classes. κ therefore provides a baseline for comparing the robustness to adversarial perturbations, and we say that f is not robust to adversarial perturbations when $\widehat{\rho}_{\text{adv}}(f) \ll \kappa$. We also compare the adversarial robustness of the classifiers with the robustness to random uniform noise. We estimate $\Delta_{\text{unif},\epsilon}(x; f)$ using a line search procedure that finds the largest η for which the condition

$$\frac{1}{J}\#\{1 \leq j \leq J : f(x + n_j)f(x) \leq 0\} \leq \epsilon,$$

is satisfied, where $n_1, \ldots, n_J$ are iid samples from the sphere $\eta\mathbb{S}$. By calling this estimate $\widehat{\Delta}_{\text{unif},\epsilon}(x; f)$, the robustness of f to uniform random noise is the empirical average over all training points $\widehat{\rho}_{\text{unif},\epsilon}(f) = \frac{1}{m}\sum_{i=1}^{m} \widehat{\Delta}_{\text{unif},\epsilon}(x_i; f)$. In the experiments, we set $J = 500$, and $\epsilon = 0.01$.[7]

(a) (b) $\Delta_{\text{adv}} = 0.08$ (c) $\Delta_{\text{adv}} = 0.19$ (d) $\Delta_{\text{adv}} = 0.21$ (e) $\Delta_{\text{adv}} = 0.15$ (f) $\Delta_{\text{adv}} = 0.41$ (g) $\Delta_{\text{unif},\epsilon} = 0.8$

Figure 6: Original image (a) and minimally perturbed images (b-f) that switch the estimated label of linear (b), quadratic (c), cubic (d), RBF(1) (e), RBF(0.1) (f) classifiers. The image in (g) corresponds to the original image perturbed with a random uniform noise of norm $\Delta_{\text{unif},\epsilon}(x; f)$, where f is the learned linear classifier. That is, the linear classifier gives the same label to (a) and (g), with high probability. The norms of the perturbations are reported in each case.

(a) (b) $\Delta_{\text{adv}} = 0.04$ (c) $\Delta_{\text{adv}} = 0.02$ (d) $\Delta_{\text{adv}} = 0.03$ (e) $\Delta_{\text{adv}} = 0.03$ (f) $\Delta_{\text{adv}} = 0.05$ (g) $\Delta_{\text{unif},\epsilon} = 0.8$

Figure 7: Same as Fig. 6, but for the "airplane" vs. "automobile" classification task.

We first consider a classification task on the MNIST handwritten digits dataset LeCun et al. (1998). We consider a digit "4" vs. digit "5" binary classification task, with $2,000$ and $1,000$ randomly chosen images for training and testing, respectively. In addition, a small random translation is applied to all images, and the images are normalized to be of unit Euclidean norm. Table 2 reports the accuracy of the different classifiers, and their robustness to adversarial and random perturbations. Despite the fact that L-SVM performs fairly well on this classification task (both on training and testing), it is highly non robust to small adversarial perturbations. Indeed, $\widehat{\rho}_{\text{adv}}(f)$ is one order of magnitude smaller than $\kappa = 0.72$. Visually, this translates to an adversarial perturbation that is hardly perceptible. The instability of the linear classifier to adversarial perturbations is not surprising, as $\frac{1}{2}\|\mathbb{E}_{\mu_1}(x) - \mathbb{E}_{\mu_{-1}}(x)\|_2$ is small (see Table 4). In addition to improving the accuracy, the more flexible classifiers are also more robust to adversarial perturbations. That is, the third order classifier is slightly more robust than the second order one, and RBF-SVM with small width $\sigma^2 = 0.1$ is more robust than with $\sigma^2 = 1$. Note that σ controls the flexibility of the classifier in a similar way to the degree in the polynomial kernel. Interestingly, in this relatively easy classification task, RBF-SVM(0.1) achieves both a good performance, and a high robustness to adversarial perturbations. Concerning the robustness to random uniform noise, the results in Table 2 confirm the large gap between adversarial and random robustness for the linear classifier, as predicted by Theorem 4.2. Moreover, the results suggest that this gap is maintained for polynomial SVM. Fig. 6 illustrates the robustness of the different classifiers on an example image.

[7] We compute the robustness to uniform random noise of all classifiers, except RBF-SVM, as this classifier is often asymmetric, assigning to one of

Model	Train error (%)	Test error (%)	$\widehat{\rho}_{\text{adv}}$	$\widehat{\rho}_{\text{unif},\epsilon}$
L-SVM	14.5	21.3	0.04	0.94
poly-SVM(2)	4.2	15.3	0.03	0.73
poly-SVM(3)	4	15	0.04	0.89
RBF-SVM(1)	7.6	16	0.04	-
RBF-SVM(0.1)	0	13.1	0.06	-

Table 3: Training and testing accuracy of different models, and robustness to adversarial noise for the CIFAR task. Note that for this example, we have $\kappa = 0.39$.

	Digits	Natural images
κ	0.72	0.39
$\|p_1\mathbb{E}_{\mu_1}(x) - p_{-1}\mathbb{E}_{\mu_{-1}}(x)\|_2$	0.14	0.06
$2\sqrt{K}\|p_1 C_1 - p_{-1}C_{-1}\|_*$	1.4	0.87

Table 4: The parameter κ, and distinguishability measures for the two classification tasks. For the numerical computation, we used $K = 1$.

We now turn to a natural image classification task, with images taken from the CIFAR-10 database Krizhevsky & Hinton (2009). The database contains 10 classes of 32×32 RGB images. We restrict the dataset to the first two classes ("airplane" and "automobile"), and consider a subset of the original data, with $1,000$ images for training, and $1,000$ for testing. Moreover, all images are normalized to be of unit Euclidean norm. Compared to the first dataset, this task is more difficult, as the variability of the images is much larger than for digits. We report the results in Table 3. It can be seen that *all* classifiers are not robust to adversarial perturbations for this experiment, as $\rho_{\text{adv}}(f) \ll \kappa = 0.39$. Despite that, all classifiers (except L-SVM) achieve an accuracy around 85%, and a training accuracy above 92%, and are robust to uniform random noise. Fig. 7 illustrates the robustness to adversarial and random noise of the learned classifiers, on an example image of the dataset. Compared to the digits dataset, the distinguishability measures for this task are smaller (see Table 4). Our theoretical analysis therefore predicts a lower limit on the adversarial robustness of linear and quadratic classifiers for this task (even though the bound for quadratic classifiers is far from the achieved robustness of poly-SVM(2) in this example).

The instability of all classifiers to adversarial perturbations on this task suggests that the essence of the classification task was not correctly captured by these classifiers, even if a fairly good test accuracy is reached. To reach better robustness, two possibilities exist: use a more flexible family of classifiers, or use a better training algorithm for the tested nonlinear classifiers. The latter solution seems possible, as the limit for quadratic classifiers suggests that there is still room to improve the robustness of these classifiers.

7 Discussion and perspectives

The existence of a limit on the adversarial robustness of classifiers is an important phenomenon with many practical implications, and opens many avenues for future research. For the family of linear classifiers, the established limit is very small for most problems of interest. Hence, linear classifiers are usually not robust to adversarial noise (even though robustness to random noise might be achieved). This is however different for nonlinear classifiers: for the family of quadratic classifiers, the limit on adversarial robustness is usually larger than for linear classifiers, which gives hope to have classifiers that are robust to adversarial perturbations. In fact, by using an appropriate training procedure, it might be possible to get closer to the theoretical bound. For general nonlinear classifiers, designing training procedures that specifically take into account the robustness in the learning is an important future work. We also believe that identifying the theoretical limit on the robustness to adversarial perturbations in terms of distinguishability measures (similar to Theorem 4.1 and 5.1) for general families of classifiers would be very interesting. In particular, identifying this limit for deep neural networks would be a great step towards having a better understanding of deep nets, and their relation with human vision.

the classes "small pockets" in the ambient space, and the rest of the space is assigned to the other class. In these cases, the robustness to uniform random noise can be equal to infinity for one of the classes, for a given ϵ.

A Proof of Theorem 4.1

Let $f(x) = w^T x + b$, such that $|b| \leq M\|w\|_2$. Our goal is to derive an upper bound on $\rho_{\text{adv}}(f) = \mathbb{E}_\mu(\Delta_{\text{adv}}(x; f)) = \frac{1}{\|w\|_2}\mathbb{E}_\mu(|f(x)|)$. We recall that μ_1 and μ_{-1} are the distributions of class 1 and class -1, respectively. We have:

$$\begin{aligned}\mathbb{E}_\mu(|f(x)|) &= p_1\mathbb{E}_{\mu_1}(|f(x)|) + p_{-1}\mathbb{E}_{\mu_{-1}}(|f(x)|)\\ &= p_1\Big(\mathbb{P}_{\mu_1}(f(x)\geq 0)\mathbb{E}_{\mu_1}(f(x)|f(x)\geq 0) - \mathbb{P}_{\mu_1}(f(x)<0)\mathbb{E}_{\mu_1}(f(x)|f(x)<0)\Big)\\ &+ p_{-1}\Big(-\mathbb{P}_{\mu_{-1}}(f(x)<0)\mathbb{E}_{\mu_{-1}}(f(x)|f(x)<0) + \mathbb{P}_{\mu_{-1}}(f(x)\geq 0)\mathbb{E}_{\mu_{-1}}(f(x)|f(x)\geq 0)\Big), \qquad (10)\end{aligned}$$

where we have conditioned successively on the events $y(x) = \pm 1$, and $f(x) \lessgtr 0$. Observe moreover that the following equality holds

$$-p_1\mathbb{P}_{\mu_1}(f(x)<0)\mathbb{E}_{\mu_1}(f(x)|f(x)<0) = 2p_1\mathbb{P}_{\mu_1}(f(x)<0)|\mathbb{E}_{\mu_1}(f(x)|f(x)<0)| + p_1\mathbb{P}_{\mu_1}(f(x)<0)\mathbb{E}_{\mu_1}(f(x)|f(x)<0).$$

By using a similar equality for $p_{-1}\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)\mathbb{E}_{\mu_{-1}}(f(x)|f(x)\geq 0)$, and plugging into Eq. (10), we obtain:

$$\begin{aligned}\mathbb{E}_\mu(|f(x)|) &= p_1\Big(\mathbb{P}_{\mu_1}(f(x)\geq 0)\mathbb{E}_{\mu_1}(f(x)|f(x)\geq 0) + \mathbb{P}_{\mu_1}(f(x)<0)\mathbb{E}_{\mu_1}(f(x)|f(x)<0)\Big)\\ &+ p_{-1}\Big(-\mathbb{P}_{\mu_{-1}}(f(x)<0)\mathbb{E}_{\mu_{-1}}(f(x)|f(x)<0) - \mathbb{P}_{\mu_{-1}}(f(x)\geq 0)\mathbb{E}_{\mu_{-1}}(f(x)|f(x)\geq 0\Big)\\ &+ 2p_1\mathbb{P}_{\mu_1}(f(x)<0)|\mathbb{E}_{\mu_1}(f(x)|f(x)<0)| + 2p_{-1}\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)|\mathbb{E}_{\mu_{-1}}(f(x)|f(x)\geq 0)|\\ &= p_1\mathbb{E}_{\mu_1}(f(x)) - p_{-1}\mathbb{E}_{\mu_{-1}}(f(x)) + 2p_1\mathbb{P}_{\mu_1}(f(x)<0)|\mathbb{E}_{\mu_1}(f(x)|f(x)<0)|\\ &+ 2p_{-1}\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)|\mathbb{E}_{\mu_{-1}}(f(x)|f(x)\geq 0)|.\end{aligned}$$

By using the fact that $f(x) = w^T x + b$, the above expression can be rewritten as follows:

$$\begin{aligned}\mathbb{E}_\mu(|f(x)|) &= w^T(p_1\mathbb{E}_{\mu_1}(x) - p_{-1}\mathbb{E}_{\mu_{-1}}(x)) + b(p_1 - p_{-1}) + 2p_1\mathbb{P}_{\mu_1}(f(x)<0)|\mathbb{E}_{\mu_1}(f(x)|f(x)<0)|\\ &+ 2p_{-1}\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)|\mathbb{E}_{\mu_{-1}}(f(x)|f(x)\geq 0)|.\end{aligned}$$

Moreover, observe that $|f(x)|$ is bounded from above as:

$$|f(x)| = |w^T x + b| \leq |w^T x| + |b| \leq 2\|w\|_2 M, \qquad (11)$$

where we have used the Cauchy-Schwarz inequality, together with the fact $|b| \leq \|w\|_2 M$. The conditional expectations $|\mathbb{E}_{\mu_{-1}}(f(x)|f(x)\geq 0)|$ and $|\mathbb{E}_{\mu_{-1}}(f(x)|f(x)<0)|$ are therefore bounded from above by $2\|w\|_2 M$. We obtain:

$$\mathbb{E}_\mu(|f(x)|) \leq w^T(p_1\mathbb{E}_{\mu_1}(x) - p_{-1}\mathbb{E}_{\mu_{-1}}(x)) + b(p_1 - p_{-1}) + 4\|w\|_2 M\big(p_1\mathbb{P}_{\mu_1}(f(x)<0) + p_{-1}\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)\big).$$

Observe that the term $p_1\mathbb{P}_{\mu_1}(f(x)<0) + p_{-1}\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)$ is equal to the *risk* of the classifier $R(f)$. Hence, we have:

$$\rho_{\text{adv}}(f) = \frac{1}{\|w\|_2}\mathbb{E}_\mu(|f(x)|) \leq \|p_1\mathbb{E}_{\mu_1}(x) - p_{-1}\mathbb{E}_{\mu_{-1}}(x)\|_2 + M|p_1 - p_{-1}| + 4MR(f),$$

where we made use once again of the Cauchy-Schwarz inequality, together with the fact that $|b| \leq M\|w\|_2$.

When $b = 0$, the inequality in (11) can be tightened and we have $|f(x)| \leq \|w\|_2 M$. Therefore, if $b = 0$ and $p_1 = p_{-1} = 1/2$, the upper bound on the adversarial robustness is

$$\rho_{\text{adv}}(f) \leq \frac{1}{2}\|\mathbb{E}_{\mu_1}(x) - \mathbb{E}_{\mu_{-1}}(x)\|_2 + 2MR(f).$$

This concludes the proof of the theorem.

B Proof of Theorem 4.2

The proof of this theorem relies on the concentration of measure on the sphere. The following result from Matoušek (2002) precisely bounds the measure of a spherical cap.

Theorem B.1. *Let $\mathscr{C}(\tau) = \{x \in \mathbb{S}^{d-1} : x_1 \geq \tau\}$ denote the spherical cap of height $1 - \tau$. Then for $0 \leq \tau \leq \sqrt{2/d}$, we have $\frac{1}{12} \leq \mathbb{P}(\mathscr{C}(\tau)) \leq \frac{1}{2}$, and for $\sqrt{2/d} \leq \tau < 1$, we have:*

$$\frac{1}{6\tau\sqrt{d}}(1-\tau^2)^{\frac{d-1}{2}} \leq \mathbb{P}(\mathscr{C}(\tau)) \leq \frac{1}{2\tau\sqrt{d}}(1-\tau^2)^{\frac{d-1}{2}}.$$

Based on Theorem B.1, we show the following result:

Lemma B.2. *Let w be a vector of unit ℓ_2 norm in $\mathbb{R}^d$. Let $\tau \in [0, 1)$, and x be a vector sampled uniformly at random from the unit sphere in $\mathbb{R}^d$. Then,*

$$\frac{1}{12}(1-\tau^2)^d \leq \mathbb{P}(\{w^T x \geq \tau\}) \leq 2\exp\left(-\frac{\tau^2 d}{2}\right).$$

Proof. Using an appropriate change of basis, we can assume that $w = (1, 0, \ldots, 0)^T$. For $\tau \in [\sqrt{2/d}, 1)$, we have

$$\mathbb{P}(\{x_1 \geq \tau\}) \overset{(a)}{\leq} \frac{1}{2\tau\sqrt{d}}(1-\tau^2)^{\frac{d-1}{2}} \overset{(b)}{\leq} 2\exp(-\tau^2 d/2),$$

where (a) uses the upper bound of Theorem B.1, and (b) uses the inequality $(1-\tau^2) \leq \exp(-\tau^2)$. Note moreover that for $\tau \in [0, \sqrt{2/d})$, the inequality $2\exp(-\tau^2 d/2) \geq 2\exp(-1) \geq 1/2$ holds, which proves the upper bound.

We now prove the lower bound. Observe that the following lower bound holds for $(\tau\sqrt{d})^{-1}$, for any $\tau \in [\sqrt{2/d}, 1)$:

$$\frac{1}{\tau\sqrt{d}} \geq \exp(-\tau^2 d/2).$$

To see this, note that the maximum of the function $a \mapsto \ln(a)/a^2$ is equal to $1/(2e) \leq 1/2$. Therefore, $\ln(\tau\sqrt{d})/(\tau^2 d) \leq 1/2$, or equivalently, $(\tau\sqrt{d})^{-1} \geq \exp(-\tau^2 d/2)$. Therefore, we get $\frac{1}{\tau\sqrt{d}} \geq (1-\tau^2)^{d/2}$, and using Theorem B.1, we obtain for any $\tau \in [\sqrt{2/d}, 1)$:

$$\mathbb{P}(\{x_1 \geq \tau\}) \geq \frac{1}{6\tau\sqrt{d}}(1-\tau^2)^{\frac{d-1}{2}} \geq \frac{1}{12}(1-\tau^2)^d.$$

Note also that this inequality holds for $\tau \in [0, \sqrt{2/d}]$, as $\frac{1}{12}(1-\tau^2)^d \leq \frac{1}{12}$. □

Armed with the concentration of measure result on the sphere, we now focus on the proof of Theorem 4.2. Let $f(x) = w^T x + b$. Let x be fixed such that $f(x) > 0$, and let $\eta > 0$ and $\epsilon \in (0, 1/12)$. Then,

$$\begin{aligned}
\mathbb{P}_{n\sim\eta\mathbb{S}}\left(f(x+n) \leq 0\right) &= \mathbb{P}_{n\sim\eta\mathbb{S}}\left(w^T n \leq -w^T x - b\right) \\
&= \mathbb{P}_{n\sim\eta\mathbb{S}}\left(w^T n/\|w\|_2 \leq -\Delta_{\text{adv}}(x; f)\right) \\
&= \mathbb{P}_{n\sim\mathbb{S}}\left(w^T n/\|w\|_2 \leq -\Delta_{\text{adv}}(x; f)/\eta\right)
\end{aligned}$$

Using the upper bound in Lemma B.2, we obtain:

$$\mathbb{P}_{n\sim\eta\mathbb{S}}\left(f(x+n) \leq 0\right) \leq 2\exp\left(-\frac{\Delta_{\text{adv}}(x; f)^2 d}{2\eta^2}\right).$$

Therefore, for $\eta = (2\ln(2/\epsilon))^{-1/2}\sqrt{d}\Delta_{\text{adv}}(x; f) = C_1(\epsilon)\sqrt{d}\Delta_{\text{adv}}(x; f)$, we obtain $\mathbb{P}_{n\sim\eta\mathbb{S}}(f(x+n) \leq 0) \leq \epsilon$, and we deduce that

$$\Delta_{\text{unif},\epsilon}(x; f) \geq C_1(\epsilon)\sqrt{d}\Delta_{\text{adv}}(x; f).$$

Using the lower bound result of Lemma B.2, we have:

$$\frac{1}{12}\left(1 - \frac{\Delta^2_{\text{adv}}(x; f)}{\eta^2}\right)^d \leq \mathbb{P}_{n\sim\eta\mathbb{S}}\left(f(x+n) \leq 0\right)$$

This implies that for any $\eta \geq \frac{\Delta_{\text{adv}}(x;f)}{\sqrt{1-(12\epsilon)^{1/d}}} = \widetilde{C_2}(\epsilon, d)\Delta_{\text{adv}}(x; f)$, we have $\mathbb{P}_{n\sim\eta\mathbb{S}}\left(f(x+n) \leq 0\right) \geq \epsilon$. Hence, we obtain the following upper bound on $\Delta_{\text{unif},\epsilon}(x; f)$:

$$\Delta_{\text{unif},\epsilon}(x; f) \leq \widetilde{C_2}(\epsilon, d)\Delta_{\text{adv}}(x; f).$$

We also derive a lower bound on $\Delta_{\text{unif},\epsilon}(x; f)$ of the form $C_2(\epsilon)\sqrt{d}\Delta_{\text{adv}}(x; f)$ by noting that

$$\widetilde{C_2}(\epsilon, d)d^{-1/2} = \frac{1}{\sqrt{d(1-(12\epsilon)^{1/d})}} \leq \frac{1}{\sqrt{1-12\epsilon}} = C_2(\epsilon),$$

where we have used the fact that $\frac{1}{\sqrt{d(1-(12\epsilon)^{1/d})}}$ is a decreasing function of d. To see that this function is indeed decreasing, note that its derivative (with respect to d) can be written as $P(d)\left(d(\bar{\epsilon}^{1/d} - 1) - \bar{\epsilon}^{1/d}\ln(\bar{\epsilon})\right)$, with $P(d)$ non-negative, and $\bar{\epsilon} = 12\epsilon$. Then, by using the inequality $\ln((1/\bar{\epsilon})^{1/d}) \leq (1/\bar{\epsilon})^{1/d} - 1$, the negativity of the derivative follows.

By combining the lower and upper bounds, and taking the expectations on both sides of the inequality, we obtain:

$$C_1(\epsilon)\sqrt{d}\mathbb{E}_\mu\left(\Delta_{\text{adv}}(x;f)1_{f(x)>0}\right) \leq \mathbb{E}_\mu\left(\Delta_{\text{unif},\epsilon}(x;f)1_{f(x)>0}\right) \leq \widetilde{C_2}(\epsilon,d)\mathbb{E}_\mu\left(\Delta_{\text{adv}}(x;f)1_{f(x)>0}\right) \\ \leq C_2(\epsilon)\sqrt{d}\mathbb{E}_\mu\left(\Delta_{\text{adv}}(x;f)1_{f(x)>0}\right).$$

A similar result can be proven for x such that $f(x) \leq 0$. We therefore conclude that

$$\max(C_1(\epsilon)\sqrt{d},1)\rho_{\text{adv}}(f) \leq \rho_{\text{unif},\epsilon}(f) \leq \widetilde{C_2}(\epsilon,d)\rho_{\text{adv}}(f) \leq C_2(\epsilon)\sqrt{d}\rho_{\text{adv}}(f),$$

where we have used the inequality $\rho_{\text{unif},\epsilon}(f) \geq \rho_{\text{adv}}(f)$.

C Proof of Theorem 5.1

In a first step of the proof, we show that for quadratic functions, the distance from a point x satisfying $f(x) \geq 0$ to the set $\{z : f(z) \leq 0\}$ is bounded by a term proportional to $\sqrt{f(x)}$.

Lemma C.1. *Consider the quadratic form $f(x) = x^T Ax$ such that $\lambda_{\min}(A) < 0$. Let x be such that $f(x) \geq 0$. Then, there exists $r \in \mathbb{R}^d$ such that $f(x+r) \leq 0$ and $\|r\|_2 \leq \sqrt{f(x)/|\lambda_{\min}(A)|}$.*

Proof. Assume without loss of generality that A is diagonal (this can be done using an appropriate change of basis). Let $\nu = -\lambda_{\min}(A)$. We have $f(x) = \sum_{i=1}^{d-1} \lambda_i x_i^2 - \nu x_d^2$. By setting $r_i = 0$ for all $i \in \{1,\dots,d-1\}$ and $r_d = \text{sign}(x_d)\sqrt{f(x)/\nu}$, (where $\text{sign}(x) = 1$ if $x \geq 0$ and -1 otherwise) we have

$$\begin{aligned} f(x+r) &= \sum_{i=1}^{d-1} \lambda_i x_i^2 - \nu(x_d + \text{sgn}(x_d)\sqrt{f(x)/\nu})^2 \\ &= f(x) - 2\nu x_d \text{sgn}(x_d)\sqrt{f(x)/\nu} - f(x) \\ &= -2\nu|x_d|\sqrt{f(x)/\nu} \leq 0, \end{aligned}$$

which concludes the proof of the lemma. □

We now prove Theorem 5.1. The goal is to upper bound $\rho_{\text{adv}}(f) = \mathbb{E}_\mu(\Delta_{\text{adv}}(x;f))$, when $f(x) = x^T Ax$. We have:

$$\begin{aligned} \rho_{\text{adv}}(f) &= p_1\mathbb{E}_{\mu_1}(\Delta_{\text{adv}}(x;f)) + p_{-1}\mathbb{E}_{\mu_{-1}}(\Delta_{\text{adv}}(x;f)) \\ &= p_1\Big(\mathbb{E}_{\mu_1}(\Delta_{\text{adv}}(x;f)|f(x)\geq 0)\mathbb{P}_{\mu_1}(f(x)\geq 0) + \mathbb{E}_{\mu_1}(\Delta_{\text{adv}}(x;f)|f(x)<0)\mathbb{P}_{\mu_1}(f(x)<0)\Big) \\ &+ p_{-1}\Big(\mathbb{E}_{\mu_{-1}}(\Delta_{\text{adv}}(x;f)|f(x)<0)\mathbb{P}_{\mu_{-1}}(f(x)<0) + \mathbb{E}_{\mu_{-1}}(\Delta_{\text{adv}}(x;f)|f(x)\geq 0)\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)\Big). \end{aligned}$$

By using Lemma C.1 successively on both functions $f(x)$ and $-f(x)$, we obtain

$$\begin{aligned} \mathbb{E}_{\mu_{\pm 1}}(\Delta_{\text{adv}}(x;f)|f(x)\geq 0) &\leq |\lambda_{\min}(A)|^{-1/2}\mathbb{E}_{\mu_{\pm 1}}(\sqrt{f(x)}|f(x)\geq 0), \\ \mathbb{E}_{\mu_{\pm 1}}(\Delta_{\text{adv}}(x;f)|f(x)<0) &\leq |\lambda_{\max}(A)|^{-1/2}\mathbb{E}_{\mu_{\pm 1}}(\sqrt{-f(x)}|f(x)<0). \end{aligned}$$

We define $\overline{\lambda} = \max(|\lambda_{\min}(A)|^{-1/2}, |\lambda_{\max}(A)|^{-1/2})$. The following inequality on $\rho_{\text{adv}}(f)$ is obtained

$$\begin{aligned} \rho_{\text{adv}}(f) \leq \overline{\lambda}\Big(&p_1\mathbb{E}_{\mu_1}(\sqrt{f(x)}|f(x)\geq 0)\mathbb{P}_{\mu_1}(f(x)\geq 0) + p_1\mathbb{E}_{\mu_1}(\sqrt{-f(x)}|f(x)<0)\mathbb{P}_{\mu_1}(f(x)<0) \\ &+p_{-1}\mathbb{E}_{\mu_{-1}}(\sqrt{-f(x)}|f(x)<0)\mathbb{P}_{\mu_{-1}}(f(x)<0) + p_{-1}\mathbb{E}_{\mu_{-1}}(\sqrt{f(x)}|f(x)\geq 0)\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)\Big). \end{aligned}$$

For any random variable X, we have $\mathbb{E}(\sqrt{X}) \leq \sqrt{\mathbb{E}(X)}$. Using this inequality, we get

$$\begin{aligned} \rho_{\text{adv}}(f) \leq \overline{\lambda}\Big(&\sqrt{p_1\mathbb{E}_{\mu_1}(f(x)|f(x)\geq 0)\mathbb{P}_{\mu_1}(f(x)\geq 0)} + \sqrt{p_1\mathbb{E}_{\mu_1}(-f(x)|f(x)<0)\mathbb{P}_{\mu_1}(f(x)<0)} \\ &+\sqrt{p_{-1}\mathbb{E}_{\mu_{-1}}(-f(x)|f(x)<0)\mathbb{P}_{\mu_{-1}}(f(x)<0)} + \sqrt{p_{-1}\mathbb{E}_{\mu_{-1}}(f(x)|f(x)\geq 0)\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)}\Big) \end{aligned} \quad (12)$$

Observe moreover that for any non-negative real numbers z_1 and z_2, we have $\sqrt{z_1}+\sqrt{z_2} \leq \sqrt{2(z_1+z_2)}$. By applying twice this inequality, we obtain $\sum_{i=1}^4 \sqrt{z_i} \leq 2\sqrt{\sum_{i=1}^4 z_i}$, for all $z_1,\dots,z_4$ in $\mathbb{R}^+$. Using this inequality in (12), we obtain

$$\begin{aligned} \rho_{\text{adv}}(f) \leq 2\overline{\lambda}\Big(&p_1\mathbb{E}_{\mu_1}(f(x)|f(x)\geq 0)\mathbb{P}_{\mu_1}(f(x)\geq 0) - p_1\mathbb{E}_{\mu_1}(f(x)|f(x)<0)\mathbb{P}_{\mu_1}(f(x)<0) \\ &-p_{-1}\mathbb{E}_{\mu_{-1}}(f(x)|f(x)<0)\mathbb{P}_{\mu_{-1}}(f(x)<0) + p_{-1}\mathbb{E}_{\mu_{-1}}(f(x)|f(x)\geq 0)\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)\Big)^{1/2} \end{aligned}$$

At this point, similarly to the linear case (Section A), we make use of the following equality

$$-p_1\mathbb{P}_{\mu_1}(f(x)<0)\mathbb{E}_{\mu_1}(f(x)|f(x)<0) = 2p_1\mathbb{P}_{\mu_1}(f(x)<0)|\mathbb{E}_{\mu_1}(f(x)|f(x)<0)| + p_1\mathbb{P}_{\mu_1}(f(x)<0)\mathbb{E}_{\mu_1}(f(x)|f(x)<0).$$

Using the above equality along with a similar one for $p_{-1}\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)\mathbb{E}_{\mu_{-1}}(f(x)|f(x)\geq 0)$, the following upper bound is obtained

$$\begin{aligned}\rho_{\text{adv}}(f) \leq 2\overline{\lambda}\Big(&p_1\mathbb{E}_{\mu_1}(f(x)|f(x)\geq 0)\mathbb{P}_{\mu_1}(f(x)\geq 0) + p_1\mathbb{E}_{\mu_1}(f(x)|f(x)<0)\mathbb{P}_{\mu_1}(f(x)<0)\\ &-p_{-1}\mathbb{E}_{\mu_{-1}}(f(x)|f(x)<0)\mathbb{P}_{\mu_{-1}}(f(x)<0) - p_{-1}\mathbb{E}_{\mu_{-1}}(f(x)|f(x)\geq 0)\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)\\ &+2p_1|\mathbb{E}_{\mu_1}(f(x)|f(x)<0)|\mathbb{P}_{\mu_1}(f(x)<0) + 2p_{-1}|\mathbb{E}_{\mu_{-1}}(f(x)|f(x)\geq 0)|\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)\Big)^{1/2},\end{aligned}$$

which simplifies to

$$\begin{aligned}\rho_{\text{adv}}(f) \leq 2\overline{\lambda}\Big(&p_1\mathbb{E}_{\mu_1}(f(x)) - p_{-1}\mathbb{E}_{\mu_{-1}}(f(x)) + 2p_1|\mathbb{E}_{\mu_1}(f(x)|f(x)<0)|\mathbb{P}_{\mu_1}(f(x)<0)\\ &+ 2p_{-1}|\mathbb{E}_{\mu_{-1}}(f(x)|f(x)\geq 0)|\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)\Big)^{1/2}\end{aligned}$$

By using the quadratic form of A, we get

$$\rho_{\text{adv}}(f) \leq 2\overline{\lambda}\Big(p_1\mathbb{E}_{\mu_1}(x^TAx) - p_{-1}\mathbb{E}_{\mu_{-1}}(x^TAx) + 2\|A\|MR(f)\Big)^{1/2},$$

where we have used $f(x) = x^TAx \leq \|A\|\|x\|_2$, with $\|A\|$ the spectral norm of A, and the fact that $R(f) = p_1\mathbb{P}_{\mu_1}(f(x) < 0) + p_{-1}\mathbb{P}_{\mu_{-1}}(f(x)\geq 0)$. We finally obtain

$$\begin{aligned}\rho_{\text{adv}}(f) &\leq 2\overline{\lambda}\Big(\sum_{i,j} a_{ij}\left(p_1\mathbb{E}_{\mu_1}(x_ix_j) - p_{-1}\mathbb{E}_{\mu_{-1}}(x_ix_j)\right) + 2\|A\|MR(f)\Big)^{1/2}\\ &\leq 2\overline{\lambda}\sqrt{\|A\|}\sqrt{\|p_1C_1 - p_{-1}C_{-1}\|_* + 2MR(f)},\end{aligned}$$

where the last inequality is obtained using the generalized Cauchy-Schwarz inequality, $\|\|_*$ denotes the nuclear norm and $C_{\pm 1}(i,j) = \mathbb{E}_{\mu_{\pm 1}}(x_ix_j)$. Note finally that since A satisfies

$$\max\left(\left|\frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}\right|, \left|\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}\right|\right) \leq K,$$

the inequality $\overline{\lambda}\sqrt{\|A\|} \leq \sqrt{K}$ holds, as $\|A\| = \max(|\lambda_{\min}(A)|, |\lambda_{\max}(A)|)$. We therefore get:

$$\rho_{\text{adv}}(f) \leq 2\sqrt{K\|p_1C_1 - p_{-1}C_{-1}\|_* + 2MKR(f)},$$

which concludes the proof.

D Vertical-horizontal example: quadratic classifier

We consider the quadratic classifier $f_{\text{quad}}(x) = x^TAx$, with

$$A = \frac{1}{2}\begin{bmatrix}0 & 1 & -1 & 0\\ 1 & 0 & 0 & -1\\ -1 & 0 & 0 & 1\\ 0 & -1 & 1 & 0\end{bmatrix}.$$

We perform a change of basis, and work in the diagonalizing basis of A, denoted by P. We have

$$P = \frac{1}{2}\begin{bmatrix}1 & 1 & -1 & -1\\ 0 & -\sqrt{2} & -\sqrt{2} & 0\\ \sqrt{2} & 0 & 0 & \sqrt{2}\\ 1 & -1 & 1 & -1\end{bmatrix},$$

$$A = P^T\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & -1\end{bmatrix}P.$$

By letting $\tilde{x} = Px$, we have:

$$f_{\text{quad}}(\tilde{x}) = \tilde{x}_1^2 - \tilde{x}_4^2.$$

Given a point x and label y, the following problem is solved to find the minimal perturbation that switches the estimated label:

$$\min_{\tilde{r}} \tilde{r}_1^2 + \tilde{r}_4^2 \text{ s.t. } y((\tilde{x}_1 + \tilde{r}_1)^2 - (\tilde{x}_4 + \tilde{r}_4)^2) \leq 0.$$

Let us consider the first datapoint $x = [1 + a, 1 + a, a, a]^T$ (the other points can be handled in an exactly similar fashion). Then, it is easy to see that $\tilde{x}_1 = 1$ and $\tilde{x}_4 = 0$, and the optimal point is achieved for $\tilde{r}_1 = -1/2$ and $\tilde{r}_4 = 1/2$. In the original space, this point corresponds to $r = P^T\tilde{r} = [0, -1/2, 1/2, 0]^T$. Therefore, $\|r\|_2 = 1/\sqrt{2}$, and we obtain $\rho_{\text{adv}}(f_{\text{quad}}) = 1/\sqrt{2}$.

Acknowledgments

The authors would like to thank Hamza Fawzi and Ian Goodfellow for fruitful discussions and comments on an early draft of the paper. We would also like to thank Guillaume Aubrun for pointing out a reference for Theorem B.1.

References

Barreno, M., Nelson, B., Sears, R., Joseph, A., and Tygar, D. Can machine learning be secure? In *ACM Symposium on Information, computer and communications security*, pp. 16–25, 2006.

Biggio, B., Nelson, B., and Laskov, P. Poisoning attacks against support vector machines. In *International Conference on Machine Learning (ICML)*, 2012.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Chalupka, K., Perona, P., and Eberhardt, F. Visual causal feature learning. *arXiv preprint arXiv:1412.2309*, 2014.

Chang, C-C and Lin, C-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

Chang, Y-W., Hsieh, C-J., Chang, K-W., Ringgaard, M., and Lin, C-J. Training and testing low-degree polynomial data mappings via linear svm. *The Journal of Machine Learning Research*, 11:1471–1490, 2010.

Dalvi, N., Domingos, P., Sanghai, S., and Verma, D. Adversarial classification. In *ACM SIGKDD*, pp. 99–108, 2004.

Fan, R-W, Chang, K-W, Hsieh, C-J, Wang, X-R, and Lin, C-J. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

Goldberg, Y. and Elhadad, M. splitsvm: fast, space-efficient, non-heuristic, polynomial kernel computation for nlp applications. In *46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pp. 237–240, 2008.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gu, S. and Rigazio, L. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lugosi, G. and Pawlak, M. On the posterior-probability estimate of the error rate of nonparametric classification rules. *IEEE Transactions on Information Theory*, 40(2):475–481, 1994.

Matoušek, Jiří. *Lectures on discrete geometry*, volume 108. Springer New York, 2002.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *arXiv preprint arXiv:1412.1897*, 2014.

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.