Understanding and Quantifying Adversarial Examples Existence in Linear Classification

XU-PENG SHI[1], A. ADAM DING[1]

[1]Department of Mathematics, Northeastern University, Boston, MA, 02115, USA
E-MAIL: {shi.xup, a.ding}@northeastern.edu

Abstract:

State-of-art deep neural networks (DNN) are vulnerable to attacks by adversarial examples: a carefully designed small perturbation to the input, that is imperceptible to human, can mislead DNN. To understand the root cause of adversarial examples, we quantify the probability of adversarial example existence for linear classifiers. Previous mathematical definition of adversarial examples only involves the overall perturbation amount, and we propose a more practical relevant definition of strong adversarial examples that separately limits the perturbation along the signal direction also. We show that linear classifiers can be made robust to strong adversarial examples attack in cases where no adversarial robust linear classifiers exist under the previous definition. The results suggest that designing general strong-adversarial-robust learning systems is feasible but only through incorporating human knowledge of the underlying classification problem.

Keywords:

(Strong) Adversarial examples; (Strong) Adversarial Robustness; Gaussian mixture model

1 Introduction

The deep neural networks (DNN) are widely used as the state-of-art machining learning classification systems due to its great performance gains in recent years. Meanwhile adversarial examples, first pointed out in [9], emerges as a novel peculiar security threat against such systems: a small perturbation that is unnoticeable to human eyes can cause the DNNs to misclassify. Various adversarial algorithms have since been developed to efficiently find adversarial examples [1, 3, 4]. Various defense methods have also been proposed to prevent adversarial example attacks: Adversarial training [3, 9]; Minmax robust training [4, 8]; Input transformation [11]. However, many of the defenses are quickly broken down by new attacking methods.

For two classes of data distributed with bounded probability densities on a compact region of a high dimensional space, [7] showed that no classifier can both have low misclassification rate and be robust to adversarial examples attack. So are we left hopeless against the threat of adversarial examples? Theoretical analysis for understanding adversarial examples is needed to address this issue. [2, 3] pointed out that susceptibility of DNN classifiers to adversarial attacks could be related to their locally linear behaviours. The existence of adversarial examples is not unique to DNN, traditional linear classifiers also have adversarial examples. In this paper, we extend the understanding of adversarial examples by quantifying the probability of their existence for a simple case of linear classifiers that performs binary classification on Gaussian mixture data.

In previous literature, a data point x is mathematically defined as having an adversarial example $x' = x + v$ when the perturbation amount $\|v\|$ is small and x' is classified differently from x. This definition does not exclude genuine signal perturbation. For example, if a dog image x is perturbed to an image x' that is classified as a cat by both human and the machine classifier, then x' should not be an adversarial example even if $\|v\| = \|x' - x\|$ is small. The proper definition needs to capture the novelty of adversarial examples attack: while a human would consider two images x' and x very similar and consider both clearly as dogs, a machine classifier misclassifies x' as a cat. While defining genuine signal perturbation for general learning problems is difficult mathematically, the signal perturbation is clear in the binary linear classification for Gaussian mixture data. We therefore propose a new definition of strong-adversarial examples that limits the perturbation amount in the signal direction separately from the limit on overall perturbation amount.

In this paper, we derive quantitative formulas for the probabilities of adversarial and strong-adversarial examples existence in the binary linear classification problem. Our quantitative analysis shows that an adversarial-robust linear classifier requires much higher signal-to-noise ratio (SNR) in data than a good performing classifier does. Therefore, in many practical applications, adversarial-robust classifiers may not be available nor are such classifiers desirable. On the contrary, useful strong-adversarial-robust linear classifiers exists at the SNR similar to that required by the existence of any useful linear classifiers, however, they require better designed training algorithms.

2 Adversarial Rates Analysis of Linear Binary Classifier on Gaussian Mixture Data

We first introduce our definitions of adversarial and strong-adversarial examples, and then we characterize their existence through defining sets. Using the defining sets, we derive explicit probability rates of (strong-)adversarial examples existence for linear classifiers on Gaussian mixture data.

2.1 Definition of Adversarial and Strong-Adversarial Examples

The classical adversarial examples are defined as follows:

Definition 1. [1] *Given a classifier C, an ε-adversarial example of a data vector x is another data vector x' such that $\|x - x'\| \leq \varepsilon$ but $C(x) \neq C(x')$.*

Without loss of generality, in this paper we focus on ℓ_2 norm perturbations. If not specified, $\|\cdot\|$ in the following refers to the ℓ_2 norm. The general ℓ_p norm ($p \geq 1$) perturbation is similar, and the results will be stated in the discussion section.

For a general machine classification problem, it is reasonable to only consider adversarial examples since the signal direction is often not easily definable mathematically. Here we consider the simple binary linear classification of Gaussian mixture data where the signal direction can be clearly distinguished. For two classes labeled '$+$' and '$-$' respectively, a linear classifier is $C(x; w, b) = \{w \cdot x + b > 0\}$ where '$\cdot$' denotes the inner product of two vectors. Here the parameters w and b are respectively the weight vector and the bias term. For the classical Gaussian mixture data problem, for each of the two classes, the d-dimensional data vector x comes from a multivariate Gaussian distribution $N(\mu_i, \sigma_i^2 I_d)$, $i =$ '$+$' or '$-$'. Notice the optimal ideal classifier here is the Bayes classifier $C(x; \mu, \bar{\mu}) = \{\mu \cdot (x - \bar{\mu}) > 0\}$[2] where $\mu = \frac{1}{2}(\mu_+ - \mu_-), \bar{\mu} = \frac{1}{2}(\mu_+ + \mu_-)$.

For this problem, the data distributions of the two classes only differ in their means μ_+ and μ_-. Thus the signal direction is $\mu_0 = \mu / \|\mu\|$. Adding $2\|\mu\|$ amount of perturbation along the signal direction changes the '$-$' class data distribution to the '$+$' class data distribution exactly, rending all classifiers unable to defend against such a perturbation.

In previous literature, the adversarial examples definition does not limit perturbation along the signal direction, therefore we propose a new definition that limits the perturbation along the signal direction separately by an amount δ, we will refer these examples as *strong-adversarial examples* .

Definition 2. *Given a classifier C, an (ε, δ)-strong-adversarial example of a data vector x is another data vector x' such that $\|x - x'\| \leq \varepsilon$ and $|(x - x') \cdot \mu_0| \leq \delta$ but $C(x) \neq C(x')$.*

To illustrate the difference between the adversarial examples and the strong-adversarial examples, we consider the following examples visualized in Figure 1. Here, Figure 1(a) shows a data vector x of dimension $d = 19 \times 19 = 361$ from the '$+$' class. To visualize, each component of the data vector is mapped onto $[0, 1]$ via function $\frac{1}{2}(\tanh \frac{2x}{3} + 1)$ and then displayed in grey scale as a 19×19 image [1].

The two means μ_+ and μ_- are chosen to be zero at every component of the vector except the component corresponding to center grid cell (shown with red boundary in Figure 1). Hence the optimal Bayes classifier identifies the image as from '$+$' (or '$-$') class when the center grid cell within the red boundary appears to be white (or black). With a perturbation amount of $\varepsilon = 0.3 \times 19 = 5.7$, Figure 1(b) shows a randomly perturbed x' which is hardly distinguishable from the first image x to the human eye. This confirms that, in defending against realistic threats, ε of magnitude $O(\sqrt{d})$ needs to be studied. (Detailed discussion of ε order is in subsection 2.3.)

For a trained support vector machine (SVM) classifier, Figure 1(c) and (d) shows two adversarial examples with the same $\varepsilon = 5.7$, but only the last one in (d) is strong-adversarial for $\delta = 1.2$. The adversarial attacks present a novel threat: a machine classifier misclassifies the perturbed data points that a human would not have noted the difference. We can see that our strong-adversarial example definition focus attention on this novel threat. In contrast, under the traditional definition, the adversarial examples include examples similar to Figure 1(c) that would indeed be classified by human into another class. We now quantitatively analyze the existence of adversarial and strong-adversarial examples.

For more general classification problems, the signal direction is harder to define. But the concept of adversarial versus strong-adversarial examples still applies. Figure 2 shows an image of '1' from the MNIST data set, and two images with added perturbations. (b) shows an adversarial example obtained by the CW-attack [1] algorithm, that is misclassified by a DNN. (c)

[1] We don't distinguish the targeted and untargeted adversarial examples here because for binary classification they are the same.

[2] Here we just use the optimal Bayes classfier for balanced case since we are focusing on the balanced case in the following text.

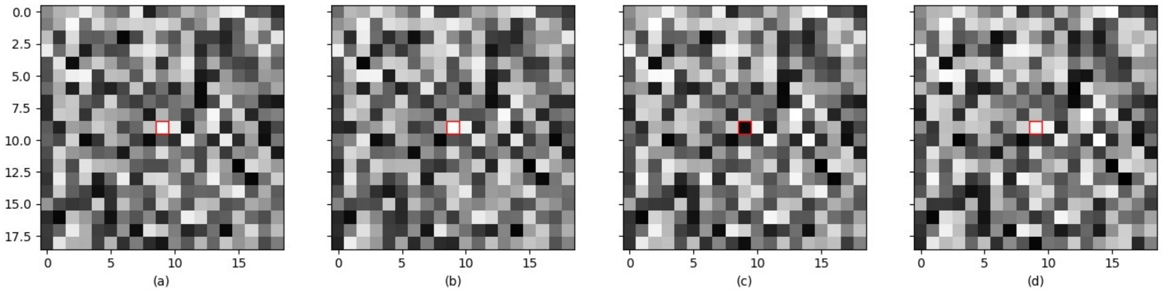

FIGURE 1. (a) a data point x from the '+' class; (b) a randomly perturbed x'; (c) an adversarial x' but not strong-adversarial; (d) a strong-adversarial x'. All three perturbations are of the same amount. The center grid cell within the red boundary contains the real class signal.

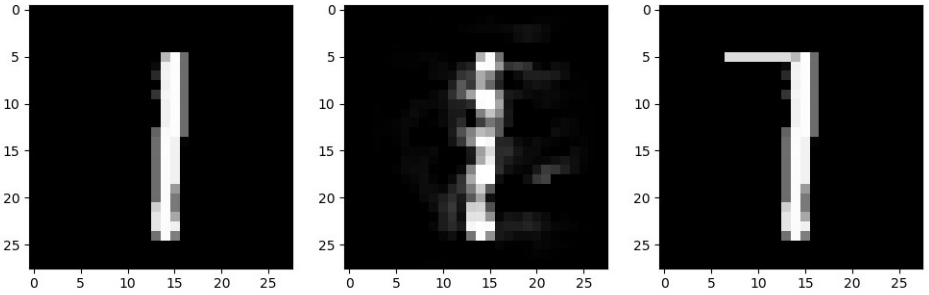

FIGURE 2. MNIST images of '1': (a) the original image, (b) an adversarial example, (c) an adversarial but not strong-adversarial example

shows an image we made with a smaller perturbation amount. If a classifier is adversarial-robust at this level, then it needs to classify both images (b) and (c) as '1'. However, classifying image (c) as '1' clearly contradicts what a human would do, rendering the usefulness of the classifier for practical applications in doubt. Generally, we should pursue a strong-adversarial-robust classifier, not an adversarial-robust one.

2.2 The Defining Sets

Here we characterize the defining sets where the (strong-) adversarial examples exist. Then we quantify the probability of data falling into these defining sets in the next subsection 2.3.

We denote $\Omega_\varepsilon = \{x : x \text{ has an } \varepsilon\text{-adversarial example}\}$ and $\Omega_{\varepsilon,\delta} = \{x : x \text{ has an } (\varepsilon,\delta)\text{-strong-adversarial example}\}$. Furthermore, for a fixed perturbation n, we denote the set where n changes classification as $\Omega(n) = \{x \in \mathbb{R}^d : C(x+n) \neq C(x)\}$.

For any data point x in Ω_ε, there exists a n with $\|n\| \leq \varepsilon$ such that $x+n$ is classified differently from x. In other words, the distance of x from the classifier's decision boundary is less than ε. For a linear classifier $C(x; w, b) = \{w \cdot x + b > 0\}$, the normal direction of its decision boundary is $n_0 = w/\|w\|$. Thus, perturbing x by ε amount along one of the two directions n_0 or $-n_0$ will cross the linear decision boundary. That is, $\Omega_\varepsilon \subseteq \Omega(\varepsilon n_0) \cup \Omega(-\varepsilon n_0)$. Since it is obvious from the definition that $\Omega_\varepsilon = \bigcup_{\|n\|\leq\varepsilon} \Omega(n) \supseteq \Omega(\varepsilon n_0) \cup \Omega(-\varepsilon n_0)$, we have $\Omega_\varepsilon = \Omega(\varepsilon n_0) \cup \Omega(-\varepsilon n_0)$. In summary, to judge if $x \in \Omega_\varepsilon$, we only need to check the perturbation along the normal direction n_0.

In contrast, our definition of strong-adversarial examples only allows δ amount of perturbation along the signal notation μ_0, hence it is not sufficient to only check perturbations εn_0 and $-\varepsilon n_0$ for judging if $x \in \Omega_{\varepsilon,\delta}$. Let θ denote the deflected angle between μ_0 and n_0. Then we can decompose n_0 into two components along and orthogonal to the signal direction μ_0 respectively. That is, $n_0 = \cos\theta\mu_0 + \sin\theta n_0$ where $n = n_0 - (n_0 \cdot \mu_0)\mu_0$ and $n_0 = n/\|n\|$. When $\varepsilon\cos\theta \leq \delta$, the adversarial example resulting from the εn_0 perturbation is also strong-adversarial by definition. When $\varepsilon\cos\theta > \delta$, however, εn_0 is no longer an allowable perturbation in the strong-adversarial example definition. Then we need to check whether classification change is caused by a perturbation of δ amount along μ_0 direction and $\sqrt{\varepsilon^2 - \delta^2}$ amount along n_0 direction. That is, to judge if $x \in \Omega_{\varepsilon,\delta}$, we need to check perturbations $u_2 = \delta\mu_0 + \sqrt{\varepsilon^2 - \delta^2} n_0$ and $-u_2$. We summarize the defining sets characterization in the following lemma.

Lemma 1. *The defining sets for ε-adversarial and (ε,δ)-strong-adversarial examples are given by:*

$$\Omega_\varepsilon = \Omega(\varepsilon n_0) \cup \Omega(-\varepsilon n_0); \quad \Omega_{\varepsilon,\delta} = \Omega(u_2) \cup \Omega(-u_2) \quad (1)$$

where $u_2 = \beta\mu_0 + \sqrt{\varepsilon^2 - \beta^2}n_0, \beta = \min(\varepsilon\cos\theta, \delta)$.

Next, we use these defining sets to quantify the probabilities of (strong-)adversarial example existence.

2.3 Adversarial and Strong-Adversarial Rates

For the binary classification problem, a random data vector comes from the Gaussian mixture distribution $p(x) = \lambda_+\varphi_+(x) + \lambda_-\varphi_-(x)$, where $\varphi_i(x)$ is the probability density function of the multivariate Gaussian $N(\mu_i, \sigma_i^2 I_d)$ and λ_i is the probability that the data vector belongs to the class of $i =$ '+' or '−'. For simplicity, we focus on the balanced classes case of $\lambda_+ = \lambda_- = 0.5$ and also $\sigma_+ = \sigma_- = \sigma$.

Adversarial Rate For a random data vector x from the '+' class, it has an ε-adversarial example x' if it is classified correctly by $w \cdot x + b > 0$ and $x \in \Omega(-\varepsilon n_0)$. Thus the adversarial rate from the '+' class is

$$\begin{aligned} &0.5pr[w \cdot x + b > 0, w \cdot (x - \varepsilon n_0) + b < 0 \,|\varphi_+(x)] \\ =&0.5pr[0 < w \cdot x + b < \varepsilon \|w\| \,|\varphi_+(x)] \end{aligned} \tag{2}$$

Under the multivariate Gaussian $N(\mu_+, \sigma^2 I_d)$ distribution $\varphi_+(x)$, $w \cdot x + b$ is univariate Gaussian with mean $w \cdot \mu_+ + b$ and variance $\|w\|^2 \sigma^2$, hence the above quantity becomes

$$0.5\left[\Phi\left(\frac{\varepsilon\|w\| - (w \cdot \mu_+ + b)}{\|w\|\sigma}\right) - \Phi\left(\frac{-(w \cdot \mu_+ + b)}{\|w\|\sigma}\right)\right] \tag{3}$$

Here $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of the standard Gaussian distribution $N(0, 1)$. Similarly, the adversarial rate from the '−' class is

$$\begin{aligned} &0.5pr[-\varepsilon\|w\| < w \cdot x + b < 0|\varphi_-(x)] \\ =&0.5\left[\Phi\left(\frac{-(w \cdot \mu_- + b)}{\|w\|\sigma}\right) - \Phi\left(-\frac{\varepsilon}{\sigma} - \frac{-w \cdot \mu_- + b}{\|w\|\sigma}\right)\right] \end{aligned} \tag{4}$$

Recall $\mu = \frac{1}{2}(\mu_+ - \mu_-), \bar{\mu} = \frac{1}{2}(\mu_+ + \mu_-)$. If we denote $b' = w \cdot \bar{\mu} + b$, then we can rewritten the expressions as $w \cdot \mu_\pm + b = \pm w \cdot \mu + b'$. Combining Eqn. (3) and (4), we have the overall adversarial rate as

$$\begin{aligned} p_{adv} =&0.5\left[\Phi\left(\frac{w \cdot \mu + b'}{\|w\|\sigma}\right) - \Phi\left(\frac{w \cdot \mu + b'}{\|w\|\sigma} - \frac{\varepsilon}{\sigma}\right)\right. \\ &\left. + \Phi\left(\frac{w \cdot \mu - b'}{\|w\|\sigma}\right) - \Phi\left(\frac{w \cdot \mu - b'}{\|w\|\sigma} - \frac{\varepsilon}{\sigma}\right)\right] \end{aligned} \tag{5}$$

Notice that the misclassification rates from the two classes are $0.5\{1 - \Phi[(w \cdot \mu + b')/(\|w\|\sigma)]\}$ and $0.5\{1 - \Phi[(w \cdot \mu - b')/(\|w\|\sigma)]\}$. Thus the overall misclassification rate is

$$p_m = 1 - 0.5\left[\Phi\left(\frac{w \cdot \mu + b'}{\|w\|\sigma}\right) + \Phi\left(\frac{w \cdot \mu - b'}{\|w\|\sigma}\right)\right] \tag{6}$$

We combine Eqn. (5) and (6) into the following Theorem.

Theorem 1. *The overall adversarial rate of a linear classifier for the balanced Gaussian mixture data is*

$$p_{adv} = 1 - p_m - 0.5\left[\Phi\left(\frac{w \cdot \mu + b'}{\|w\|\sigma} - \frac{\varepsilon}{\sigma}\right) + \Phi\left(\frac{w \cdot \mu - b'}{\|w\|\sigma} - \frac{\varepsilon}{\sigma}\right)\right] \tag{7}$$

To be robust against adversarial attacks, a linear classifier needs a low adversarial rate. For the classifier to be useful, it also needs a low misclassification rate. Hence we should look at the sum of misclassification rate and adversarial rate, which we call the *adversarial-error* rate:

$$\begin{aligned} p_{err} &= p_{adv} + p_m \\ &= 1 - 0.5\left[\Phi\left(\frac{w \cdot \mu + b'}{\|w\|\sigma} - \frac{\varepsilon}{\sigma}\right) + \Phi\left(\frac{w \cdot \mu - b'}{\|w\|\sigma} - \frac{\varepsilon}{\sigma}\right)\right] \end{aligned} \tag{8}$$

Comparing Eqn. (8) with (6), we can see why adversarial-robustness is hard to achieve. Firstly, the misclassification rate Eqn. p_m in (6) is minimized by the Bayes classifier with $b' = 0$ and $w \cdot \mu = \|w\| \|\mu\|$. Hence the best p_m value is $1 - \Phi(\|\mu\|/\sigma)$. There exists useful classifiers when $\|\mu\|/\sigma$ is big enough to make $1 - \Phi(\|\mu\|/\sigma)$ small. This is achieved for $\|\mu\|/\sigma = O(1)$. For example, when $\|\mu\|/\sigma = 3$, the misclassification rate of the Bayes classifier is around 0.1%.

However, to achieve a low adversarial-error rate in Eqn. (8), the required SNR $\|\mu\|/\sigma$ can be much bigger. When $w \cdot \mu > \varepsilon\|w\|$, a lower bound for the adversarial-error rate is

$$p_{err} \geq 1 - \Phi\left(\frac{w \cdot \mu}{\|w\|\sigma} - \frac{\varepsilon}{\sigma}\right) \geq 1 - \Phi\left(\frac{\|\mu\|}{\sigma} - \frac{\varepsilon}{\sigma}\right) \tag{9}$$

Therefore, the existence of a useful adversarial-robust linear classifier requires $\|\mu\|/\sigma - \varepsilon/\sigma = O(1)$ instead. Notice that, for this Gaussian mixture data setup, the noise in each class follows the $N(0, \sigma^2 I_d)$ distribution with an expected ℓ_2 norm square of $d\sigma^2$. Therefore, for a positive constant $\eta_a < 1$, the perturbation amount of $\varepsilon = \eta_a\sqrt{d}\sigma$ is smaller than the average noise in data and generally is hard to detect. Hence, for the typical high-dimensional data applications, an adversarial-robust linear classifier needs to protect against perturbation amount of $\varepsilon = O(\sqrt{d})$ which implies that $\|\mu\|/\sigma = O(\sqrt{d})$ is needed from Eqn. (9). Next, we show that this high SNR requirement is not needed for a strong-adversarial-robust linear classifier.

Strong-Adversarial Rate The derivation of the strong-adversarial rate is very similar to that of the adversarial rate. From Eqn. (1), the difference between the adversarial defining set and the strong-adversarial defining set is only that εn_0 is replaced by $u_2 = \beta\mu_0 + \sqrt{\varepsilon^2 - \beta^2} n_0$. Hence the strong-adversarial rate from the '$+$' class is

$$0.5pr[0 < w \cdot x + b < w \cdot u_2 | \varphi_+(x)].$$

Since $w \cdot \mu_0 = \|w\| \cos\theta$ and $w \cdot n_0 = \|w\| \sin\theta$, we have $w \cdot u_2 = (\beta \cos\theta + \sqrt{\varepsilon^2 - \beta^2} \sin\theta) \|w\|$ where $\beta = \min(\varepsilon \cos\theta, \delta)$. We denote

$$g(\varepsilon, \delta, \theta) = \beta \cos\theta + \sqrt{\varepsilon^2 - \beta^2} \sin\theta \tag{10}$$

Thus replacing $\varepsilon \|w\|$ by $g(\varepsilon, \delta, \theta) \|w\|$ in Eqn. from (3) to (8), we have the following Theorem.

Theorem 2. *The overall strong-adversarial-error rate of a linear classifier is*

$$p_{s-err} = p_{s-adv} + p_m = 1 - 0.5 \left[\Phi\left(\frac{w \cdot \mu + b'}{\|w\| \sigma} - \frac{g(\varepsilon, \delta, \theta)}{\sigma} \right) + \Phi\left(\frac{w \cdot \mu - b'}{\|w\| \sigma} - \frac{g(\varepsilon, \delta, \theta)}{\sigma} \right) \right] \tag{11}$$

Compared to the analysis above, the existence of a useful strong-adversarial-robust linear classifier requires $\|\mu\| / \sigma - g(\varepsilon, \delta, \theta)/\sigma = O(1)$ instead. Besides the overall perturbation amount ε, the function $g(\varepsilon, \delta, \theta)$ in Eqn. (10) is also affected by two other factors: the signal direction perturbation amount δ and the angle θ between the classifier and the ideal Bayes classifier. What is the practical relevant amount δ we should study? Let $\delta = \eta_s \mu = \eta_s \|\mu\|$. When $\eta_s > 1$, a δ amount perturbation along the signal direction to all '$+$' class data points will make more than half of them be classified as '$-$' by the Bayes classifier (also to human eye, e.g., Figure 1(c)). Therefore, when studying real strong-adversarial perturbations (imperceptible to human but confuses machine) mathematically, we need to focus on $\eta_s < 1$. That is, $\delta = O(1)$. Compared to the overall perturbation amount $\varepsilon = O(\sqrt{d})$ discussed earlier, we see that $\delta \ll \varepsilon$ for typical high-dimensional data applications. When $\delta \ll \varepsilon$, $g(\varepsilon, \delta, \theta) \approx \delta \cos\theta + \varepsilon \sin\theta$. Hence if the linear classifier is well-trained to have small θ and small bias b' (i.e., very close to the Bayes classifier), then its strong-adversarial-error rate is approximately $1 - \Phi[(1 - \eta_s)\|\mu\|/\sigma]$, which can be made small when SNR $\|\mu\| / \sigma$ is of order $O(1)$. That is, with good training, we can find a useful strong-adversarial-robust linear classifier when $\|\mu\| / \sigma = O(1)$. In contrast, no training can make the linear classifier to be useful and adversarial-robust unless the SNR $\|\mu\| / \sigma$ is much bigger, at the order of $O(\sqrt{d})$.

The conclusion for the analysis using ℓ_p norm is similar. One can apply the similar analysis method and show that there exists a useful strong-adversarial-robust linear classifier for constant order SNR $\|\mu\| / \sigma = O(1)$, but a useful ℓ_p-adversarial-robust linear classifier only exists when SNR is much bigger, at the order of $O(d^{min(1/p, 1/2)})$.

3 Discussions and Conclusions

In this paper, we provide clear definitions of adversarial and strong adversarial examples in the linear classification setting. Quantitative analysis shows that adversarial examples are hard to avoid but also should not be of concern in practice. Rather, we should focus on finding strong-adversarial-robust classifiers. We now consider the implications of these results on studying adversarial examples for general classifiers, and their relationship to some recent works in literature.

Shafahi et al. [7] shows that no classifier can achieve low misclassification rate and also be adversarial-robust for data distributions with bounded density on a compact region in a high-dimensional space. Our analysis does not match exactly with their impossibility statement because we are studying the Gaussian mixture case, which has positive density on the whole space. However, in spirit our results have similar implications: for the usual SNR $O(1)$ that allows low misclassification rate, generally it is impossible to be also adversarial-robust (for which a much bigger SNR $O(\sqrt{d})$ is required).

Our results, however, do show that there can be adversarial-robust classifiers under the traditional definition when the SNR is very big. Schmidt et al. [5] has also shown that, for Gaussian mixture classification problem and a particular training method, the adversarial-robustness is achievable but requires more training data than simply achieving the low misclassification rate only. Our formula indicates that useful adversarial-robust classifier do exist at the SNR level they assumed. Our study is more focused on the fundamental issue of when useful adversarial-robust classifiers exist, not which training method and what data complexity will find such a classifier. However, our formulas do indicate that an adversarial-robust classifier has to satisfy a stricter requirement than a good performing classifier. Thus either a better training method or a higher data complexity is needed for finding a useful adversarial-robust classifier, agreeing with the general theme of Schmidt et al. [5].

Our results on the existence of adversarial examples do not change qualitatively when using other ℓ_p norm to measure the perturbation: under traditional definition, useful adversarial-robust classifier exists only when the data distribution has a very big SNR of $O(d^{min(1/p,1/2)})$. For many applications where good classifiers exists (SNR of only $O(1)$), we can not pursue adversarial-robust classifier under the traditional definition 1. The current defense strategies based on such adversarial example definition likely will still be suspect to more sophisticated adversarial attacks. For certifiable adversarial-robust classifiers [4, 8], the robustness is achieved only for the perturbation amount ε high enough so that they differ from human in classifying images like those in Figure 1(c) and Figure 2(c). Thus a paradigm change is needed: we should train a classifier to be strong-adversarial-robust rather than adversarial-robust.

While the signal direction is obvious in the linear classification, the signal direction and the definition of strong-adversarial examples in general classification warrants further study. The signal direction in the linear classification here is the direction where the likelihood ratio of the two classes changes most rapidly. One reasonable extension is to define the signal direction at any data vector x as the gradient direction of the likelihood ratio at x. Then similar to definition 2, the strong-adversarial example for general classifier also restrict the change along this signal direction to the amount δ. The strong-adversarial-robust classifiers therefore are likely to be very close to the Bayes classifier. Some recent works have attempted training DNN to be close to the Bayes classifier: Wang et al. [10] uses a nearest neighbors method, and Schott et al. [6] applies the generative model techniques. In particular, Schott et al. [6] applied their method on MNIST dataset, and when applying a specifically designed attack on such a trained DNN, the adversarial examples found are semantically meaningful for humans. That is, these adversarial examples are adversarial in traditional definition but likely not strong-adversarial. The new strong-adversarial examples framework can allow theoretical quantification of the robustness for these training methods. The analysis of strong-adversarial-robustness for general classifiers such as DNN can provide a new research direction on how to defend against realistic adversarial attacks.

Acknowledgements

The research of Aidong Adam Ding is supported by National Science Foundation grant CNS-2212010.

References

[1] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, May 2017. doi: 10.1109/SP.2017.49.

[2] A. Fawzi, O. Fawzi, and P. Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, Mar 2018. ISSN 1573-0565. doi: 10.1007/s10994-017-5663-3. URL https://doi.org/10.1007/s10994-017-5663-3.

[3] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6572.

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

[5] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018.

[6] L. Schott, J. Rauber, M. Bethge, and W. Brendel. Towards the first adversarially robust neural network model on mnist. In *Seventh International Conference on Learning Representations (ICLR 2019)*, pages 1–16, 2019.

[7] A. Shafahi, W. R. Huang, C. Studer, S. Feizi, and T. Goldstein. Are adversarial examples inevitable? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1lWUoA9FQ.

[8] A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk6kPgZA-.

[9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2014.

[10] Y. Wang, S. Jha, and K. Chaudhuri. Analyzing the robustness of nearest neighbors to adversarial examples. In *International Conference on Machine Learning*, pages 5120–5129, 2018.

[11] W. Xu, D. Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.