Rademacher Complexity for Adversarially Robust Generalization [*]

Sourya Basu
Course Project: ECE 543 - Statistical Learning Theory
Course instructor: Prof. Bruce Hajeck

1 Introduction

Machine learning algorithms have shown incredible performance in several inference tasks such as image classification, speech recognition, and game playing [2–4]. Although these algorithms work excellently in natural settings (i.e., test data generated in the same way as the train data), their performance drop significantly when the test data is corrupted adversarially, at times in a manner that is imperceptible by human beings [5]. Such examples where the input to a trained machine learning model is perturbed so as to increase the probability of wrong inference is called *adversarial examples.* In image classification, adversarial examples might include something as simple as adding small perturbations imperceptible to humans, changing surrounding areas of the main object in an image or even simple rotations or translations [6,7]. The existence of such adversarial examples is alarming in several situations, because of the use machine learning algorithms in several critical examples, such as medical diagnosis or self-driving cars, e.g. self-driving cars that rely on real-time image recognition might end up with making wrong predictions in the presence of, say, different lighting or weather than the one it is trained in, or medical diagnosis might end up predicting some hence making it important for us to analyze the performance of machine learning algorithms in the presence of adversaries. The existence of such examples raise serious questions about the robustness of the existing state-of-the-art machine learning algorithm and risk of using them in critical applications.

The existence of such adversarial examples have motivated researchers to come up with algorithms that can defend against particular examples effectively, however, this also resulted in creation of more and more of such adversarial examples which the existing algorithms does not perform well on, to the extent that there is a virtual race between designing adversarial examples and designing algorithms that can defend them. As reported in the paper [1], in the current scenario, the attackers generating adversarial examples are winning, e.g. it has been shown in [8], that carefully designed gradient-based algorithms may fool most of the existing defense algorithms.

Hence, instead of designing defense algorithms against particular adversarial examples, this paper [1] takes a different approach and establishes theoretical guarantees on the performance of learning algorithms in the presence of adversaries during the testing phase. The paper [1] uses adversarial training as it appears to be quite effective against adversarial examples as described in literature [10,11] which optimizes over adversarial loss during the training phase as described next.

Let $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ be the data points drawn according to some unknown distribution $\mathcal{D}$. Let $\mathcal{F}$ be a hypothesis class and $l(f(\mathbf{x}), y)$ be the loss associated with $f \in \mathcal{F}$. This paper [1] considers l_∞ adversarial attacks wherein, an ϵ-attack can be described as follows: the adversary is allowed to observe the trained model, and given a test data, x, the adversary finds a data point x' such that $\|x - x'\|_\infty \leq \epsilon$ and $l(f(x'), y)$ is maximized. Thus, to have better test performance, the learning algorithm should solve the following optimization problem, called as the *adversarial risk minimization problem*:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \max_{\|x_i' - x_i\| \leq \epsilon} l(f(x_i'), y_i), \tag{1}$$

[*] This is a review of [1] used as a report for the course project ECE 543: Statistical learning theory.

where $\{\mathbf{x}_i, y_i\}_{i=1}^n$ are i.i.d. training examples drawn according to $\mathcal{D}$. Two main interest of researchers related to adversarial risk minimization are the following: 1) solve the (1), and 2) characterize the generalization property of the adversarial risk, i.e., the gap between the empirical risk in (1) and the expected risk $\mathbb{E}_{\mathcal{D}}[\max_{\|x-x'\|_\infty \leq \epsilon} l(f(x'), y)]$. For neural networks both these questions are still open. Further, it has also been shown, that for neural networks empirical risk minimization might not imply generalization.

This paper [1] aims at better understanding of the generalization ability of several classifiers and focus on l_∞ adversarial attacks by finding tight bounds on Rademacher complexity for such class of classifiers. The main contributions of this paper [1] can be summarized as follows: 1) they provide a tight bound on the adversarial Rademacher complexity and show that it is always greater than or equal to the Rademacher complexity in the natural setting. It is shown that when the weight vector of the classifier has a bounded l_p norm, then there is a polynomial dependency on the dimension of the data. This polynomial dependency is unavoidable, uunless, $p = 1$. 2) They provide a lower bound for Rademacher complexity for neural networks which is shown to have a dependency on the dimension of the data. Next, we discuss the problem statement and preliminary results.

2 Notations & problem setup

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y}$ be the feature and label spaces, respectively, distributed according to $\mathcal{D}$. Let $\mathcal{F} \subseteq \mathcal{V}^{\mathcal{X}}$ be the hypothesis class, where $\mathcal{V}$ could be different from the space $\mathcal{Y}$. Let $l : \mathcal{V} \times \mathcal{Y} \mapsto [0, B]$ be the loss function, where B is a positive constant. Further, $l_{\mathcal{F}} := \{(\mathbf{x}, y) \mapsto (l(f(\mathbf{x}), y)) : f \in \mathcal{F}\}$. Let $\{\mathbf{x}_i, y_i\}_{i=1}^\infty$ be n i.i.d. training examples drawn from $\mathcal{D}$, the empirical risk, $R_n(f)$, and average risk, $R(f)$ are defined using their standard definition as, $R_n(f) = \frac{1}{n}\sum_{i=1}^n l(f(\mathbf{x}_i), y_i)$ and $R(f) = E_{\mathcal{D}}[l(f(\mathbf{x}), y)]$. Also, as we know, the Rademacher complexity for a given sample $\mathcal{S} = \{z_1, z_2, \ldots, z_n\}$ is defined as follows:

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{H}) = \frac{1}{n} E_\sigma[\sup_{h \in \mathcal{H}} \sum_{i=1}^n \sigma_i h_i]$$

where σ_i are the Rademacher random variables. Let, $\mathbb{B}_x^\infty(\epsilon) := \{x' \in \mathbb{R}^d : \|x - x'\|_\infty \leq \epsilon\}$ be the ϵ ball around the data point $\mathbf{x}$, then we define the average adversarial risk as, $\tilde{R}(f) = E_{\mathcal{D}}[\max_{\mathbb{B}_x^\infty(\epsilon)} l(f(\mathbf{x}'), y)]$. Similarly, as described before, the empirical adversarial risk is denote by $\tilde{R}_n(f) = \frac{1}{n}\sum_{i=1}^n \max_{\mathbb{B}_{x_i}^\infty(\epsilon)} l(f(\mathbf{x}_i'), y_i)$. In the next section, the paper considers the linear classification case and finds tight bounds for the Rademacher complexity in the adversarial case.

3 Binary linear classifier

In this section, we discuss the results on binary linear classifiers. Let $\mathcal{Y} = \{-1, +1\}$ and let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a set of linear functions of $\mathbf{x} \in \mathcal{X}$. More specifically, $f_w(x) = \langle w, x \rangle$ and $\mathcal{F}$ consists of all such linear functions with bounded l_p norm, that is,

$$\mathcal{F} = \{f_w(x) : \|w\|_p \leq W\},$$

for some constant W. Since the linear classifiers predict the sign of the linear functions $f_w(x)$, thus we assume that $l(f_w(x), y)$ can be written as $\phi(y\langle w, x \rangle)$ where $\phi(.)$ is a monotonically non-increasing and L_ϕ-Lipschitz continuous function.

For the adversarial setting, we have,

$$\tilde{l}(f_w(x), y) = \max_{x' \in \mathbb{B}_x^\infty} l(f_w(x'), y) = \phi(min_{x' \in \mathbb{B}_x^\infty} y\langle w, x' \rangle)$$

.

So, in a way, the above equation helps us convert the expression with adversarial loss and linear function into a function with $\phi(.)$ and an adversarial function. Let us define this function class, $\tilde{F} \subseteq \mathbb{R}^{\mathcal{X}}$.

$$\tilde{F} = \{\min_{x' \in \mathbb{B}_x^\infty} y\langle w, x' \rangle : \|w\|_p \leq W\}. \tag{2}$$

From contraction inequality [12], we have $\mathfrak{R}_{\mathcal{S}}(\tilde{l}_{\mathcal{F}}) \leq L_\phi \mathfrak{R}_{\mathcal{S}}(\tilde{\mathcal{F}})$ using the Lipscitz continuity of the loss function. Next, we present the main theorem of the paper.

Theorem 1. *Let $\tilde{\mathcal{F}} := \{min_{x' \in \mathbb{B}_x^\infty(\epsilon)} y\langle w, x'\rangle : \|w\|_p \leq W\}$. Suppose that $\frac{1}{p} + \frac{1}{q} = 1$. Then there exists a universal constant $c \in (0,1)$ such that*

$$\max\{\mathfrak{R}_S(\mathcal{F}), c\epsilon W \frac{d^{\frac{1}{q}}}{\sqrt{n}}\} \leq \mathfrak{R}_S(\tilde{\mathcal{F}}) \leq \mathfrak{R}_S(\mathcal{F}) + \epsilon W \frac{d^{\frac{1}{q}}}{\sqrt{n}}$$

which implies

$$c(\mathfrak{R}_S(\mathcal{F}) + \epsilon W \frac{d^{\frac{1}{q}}}{\sqrt{n}}) \leq \mathfrak{R}_S(\tilde{\mathcal{F}}) \leq \mathfrak{R}_S(\mathcal{F}) + \epsilon W \frac{d^{\frac{1}{q}}}{\sqrt{n}}$$

Proof. By the definition of Rademacher complexity and using Holder's inequality, we have,

$\mathfrak{R}_{\mathcal{S}}(\mathcal{F}) := \frac{1}{n} E_\sigma[\sup_{\|w\|_p \leq W} \sum_{i=1}^n \sigma_i \langle w, x_i\rangle] = \frac{W}{n} E_\sigma[\| \sum_{i=1}^n \sigma_i x_i \|_q]$. Next, we will see that $\tilde{f}_w = \langle w, x\rangle - \epsilon\|w\|_1$. Note that, if $y = 1$, $\tilde{f}_w(x,y) = \min_{x' \in B_x^\infty} \langle w, x'\rangle$, else if $y = -1$, $\tilde{f}_w(x,y) = -\max_{x' \in B_x^\infty} \langle w, x'\rangle$. Thus, for $y = 1$, we have $\tilde{f}_w(x,y) = \min_{x' \in B_x^\infty} \sum_{i=1}^d w_i x'_i = \sum_{i=1}^d w_i(x_i - \epsilon sgn(w_i)) = \langle w, x\rangle - \epsilon\|w\|_1$. Similarly, when $y = -1$, we have $\tilde{f}_w(x,y) = -\langle w, x\rangle - \epsilon\|w\|_1$. Hence, we have $\tilde{f}_w = \langle w, x\rangle - \epsilon\|w\|_1$.

Thus, we have,

$$\mathfrak{R}_{\mathcal{S}}(\tilde{F}) = \frac{1}{n} E_\sigma[\sup_{\|w\|_p \leq W} \sum_{i=1}^n \sigma_i(y_i\langle w, x_i\rangle - \epsilon\|w\|_1)].$$

Now, define $u := \sum_{i=1}^n \sigma_i y_i x_i$ and $v := \epsilon \sum_{i=1}^n \sigma_i$. Thus, we have $\mathfrak{R}_{\mathcal{S}}(\tilde{F}) = \frac{1}{n} E_\sigma[\sup_{\|w\|_p \leq W} \sum_{i=1}^n \langle w, u\rangle - v\|w\|_1]$. Further, note that supremum of $\langle w, u\rangle - v\|w\|_1$ over w can only be achieved when $sgn(w_i) = sgn(u_i)$. using this and some algebraic manipulation gives,

$$\begin{aligned}\mathfrak{R}_{\mathcal{S}}(\tilde{F}) &= \frac{W}{n} E_\sigma[\|u - vsgn(u)\|_q] \\ &= \frac{W}{n} E_\sigma[\| \sum_{i=1}^n \sigma_i y_i x_i - (\epsilon \sum_{i=1}^n \sigma_i) sgn(\sum_{i=1}^n \sigma_i y_i x_i)\|_q]\end{aligned}$$

Using triangle inequality,

$$\begin{aligned}\mathfrak{R}_{\mathcal{S}}(\tilde{F}) &\leq \frac{W}{n} E_\sigma[\| \sum_{i=1}^n \sigma_i y_i x_i \|_q] - \epsilon \frac{W}{n} E_\sigma[\| \sum_{i=1}^n \sigma_i) sgn(\sum_{i=1}^n \sigma_i y_i x_i)\|_q] \\ &= \mathfrak{R}_{\mathcal{S}}(F) + \epsilon d^{\frac{1}{q}} \frac{W}{n} E_\sigma[| \sum_{i=1}^n \sigma_i)|] \\ &\leq \mathfrak{R}_{\mathcal{S}}(F) + \epsilon d^{\frac{1}{q}} \frac{W}{\sqrt{n}},\end{aligned}$$

using Khintchine's inequality. This gives the upper bound.

Next, we will find the lower bound. From previous equations and symmetry of Rademacher random variables, we have,

$$\begin{aligned}\mathfrak{R}_{\mathcal{S}}(\tilde{F}) &= \frac{W}{n} E_\sigma[\| \sum_{i=1}^n \sigma_i y_i x_i + (\epsilon \sum_{i=1}^n \sigma_i) sgn(\sum_{i=1}^n \sigma_i y_i x_i)\|_q] \\ &= \frac{W}{2n} E_\sigma[\| \sum_{i=1}^n \sigma_i y_i x_i - (\epsilon \sum_{i=1}^n \sigma_i) sgn(\sum_{i=1}^n \sigma_i y_i x_i)\|_q] + \frac{W}{2n} E_\sigma[\| \sum_{i=1}^n \sigma_i y_i x_i + (\epsilon \sum_{i=1}^n \sigma_i) sgn(\sum_{i=1}^n \sigma_i y_i x_i)\|_q] \\ &\geq \frac{W}{n} E_\sigma[\| \sum_{i=1}^n \sigma_i y_i x_i \|_q] = \mathfrak{R}_{\mathcal{S}}(F),\end{aligned}$$

where the last inequality was using triangle inequality. Thus, we have $\mathfrak{R}_{\mathcal{S}}(\tilde{F}) \geq \mathfrak{R}_{\mathcal{S}}(F)$. Similarly, we have,

$$\begin{aligned} \mathfrak{R}_{\mathcal{S}}(\tilde{F}) &= \frac{W}{n} E_\sigma[\|(\epsilon \sum_{i=1}^{n} \sigma_i) sgn(\sum_{i=1}^{n} \sigma_i y_i x_i)\|_q] \\ &= \epsilon d^{\frac{1}{q}} \frac{W}{n} E_\sigma[|\sum_{i=1}^{n} \sigma_i|] \\ &\leq c\epsilon d^{\frac{1}{q}} \frac{W}{\sqrt{n}} \text{ (for some constant c>0) }, \end{aligned}$$

where the last inequality follows from Khintchine's inequality. □

Next, we discuss the implications of the above theorem.

There are two interesting implications that one can draw from the above result. Firstly, this theorem gives a tight bound to the adversarial Rademacher complexity upto a factor of a constant. Secondly, and more importantly, this result shows the dependency of the adversarial Rademacher complexity on the dimension d of the data, i.e. the adversarial Rademacher complexity can be larger than the Rademacher complexity in natural setting by order of magnitude ($O(d^{\frac{1}{q}})$). It is also interesting to note that, for $p = 1$, $q \to \infty$, and hence $d^{\frac{1}{q}} \to 1$, and hence the dependency on d vanishes indicating that l_1 norm bound on the weight matrix might help in adversarial generalization. It is also worth noting that there is a similar result in recent literature: Cullina et al. that shows that VC-dimension for halfspace classifiers in the presence of a sample-wise norm-constrained adversary is the same as the standard VC dimension, i.e. $= d + 1$. We do not compare the two results in details due to difference in assumptions in the two work and further it is noted in the paper [1] that, the results of Cullina et. al. [13] does not provide explanation to the empirical observation that adversarially robust generalization may be hard.

A similar result is also provided in [1] for multi-class linear classifiers, which we will not discuss here since the result provided is similar to the result for the binary case. However, it might be of interest to note the implications of the result on multi-class classifiers. The theorem for multi-class classifiers imply a similar dependence of Rademacher complexity on dimension of the data which goes away if we take l_1 norm bound on the weight matrix.

4 Neural Networks

In this section feedforward neural network with ReLU activation function is considered. Each of the function in the hypothesis class corresponding to a feedforward neural network with L layers of neurons can be represented using L weight matrices, $W = (W_1, \ldots, W_L)$ as

$$f_W(x) = W_L(\rho(W_{L-1}\rho(\ldots W_1(x)))),$$

where ρ is the ReLU function defined as $\rho(t) = \max t, 0$. In this report, let us only consider the binary classification problem, hence we take $\mathcal{Y} = \{-1, +1\}$, and assume that the loss function can be written as

$$l(f_W(x), y) = \phi(y f_W(x)),$$

where $\phi(\cdot) : \mathbb{R} \mapsto [0, B]$ is a monotonically non-increasing and L_ϕ Lipschitz continuous function.

Since this is a binary classification problem as in the previous section, we have the loss function as,

$$\tilde{l}(f_W(x), y) = \max_{x' \in \mathbb{B}_x^\infty} l(f_W(x'), y) = \phi(min_{x' \in \mathbb{B}_x^\infty} y \langle W, x' \rangle).$$

The paper considers the following function class:

$$\tilde{F} = \{(\mathbf{x}, y) \mapsto \min_{x' \in B_x^\infty(\epsilon)} y f_W(x') : W = (W_1, W_2, \ldots, W_L), \prod_{h=1}^{L} \|W_h\|_\sigma \leq r\} \subseteq \mathbb{R}^{\mathcal{X} \times \{-1,1\}},$$

where $\|\cdot\|_\sigma$ is the spectral norm of a matrix. Next, we discuss the theorem providing the bound on the Rademacher complexity for feedforward neural networks.

Theorem 2. *Let* $\tilde{\mathcal{F}} := \{(x,y) \mapsto min_{x' \in \mathbb{B}_x^\infty(\epsilon)} y f_W(x') : W = (W_1, \dots, W_L), \prod_{h=1}^L \|W_h\|_\sigma \leq r\}$*. Then, there exists a universal constant* $c > 0$ *such that,*

$$\mathfrak{R}_S(\tilde{\mathcal{F}}) \geq cr(\frac{1}{n}\|X\|_F + \epsilon\sqrt{\frac{d}{n}}).$$

where, $d = \max_{h \in [L]} d_h$

The proof of this theorem follows from two different results, one of them being Thm. 1, and the other one is from a paper by Bartlett et. al. [14]. We will first state the result from [14], followed by the outline of the proof to Thm. 2.

Lemma 1. *Define the function class,*

$$\hat{\mathcal{F}} = \{\boldsymbol{x} \mapsto f_W(x) : W = (W_1, W_2, \dots, W_L), \prod_{h=1}^{L} \|W_h\|_\sigma \leq r\},$$

and $\hat{\mathcal{F}'} = \{x \mapsto \langle W, x \rangle : \|w\|_2 \leq \frac{r}{2}\}$*. Then, we have* $\hat{\mathcal{F}'} \subseteq \hat{\mathcal{F}}$*, and thus there exists a universal constant* $c > 0$ *such that,*

$$\mathfrak{R}_{\mathcal{S}}(\hat{\mathcal{F}}) \geq \frac{cr}{n}\|\boldsymbol{X}\|_F$$

Thus, by defining $\tilde{\mathcal{F}'} = \{x \mapsto min_{x' \in B_x^\infty(\epsilon)} y\langle W, x' \rangle : \|w\|_2 \leq \frac{r}{2}\} \subseteq \mathbb{R}^{\mathcal{X} \times \{-1,+1\}}$, we have $\tilde{\mathcal{F}'} \subseteq \tilde{\mathcal{F}}$. Thus, by Lemma. 1 and Thm. 1 (using $p = 2$) we have,

$$\mathfrak{R}_{\mathcal{S}}(\tilde{F}) \geq \mathfrak{R}_{\mathcal{S}}(\tilde{F'}) \geq cr(\frac{1}{n}\|X\|_F + \epsilon\sqrt{\frac{d}{n}}).$$

That completes the proof. Note that, similar to the case of linear classifiers, we have a dependency on the dimension d of the data. However, the paper shows that using some results from optimization [15]), one can design a surrogate loss function, for which this dependency on dimension can be get rid of under l_1 norm bound on the weight matrix for a single layered neural network with ReLU activation function. One may refer to [1] for further details.

Next, we look at some experimental results supporting the theoretical findings.

5 Experiments

In this section we discuss the experimental results provided in [1] for linear classifiers as well as neural networks which essentially show the dependence of generalization error on the dimension of the data, for all cases other than when weight matrix has a l_1 norm bound.

5.1 Linear classifiers

For linear classifiers, the following two implications of Thm. 1 are validated: 1) imposing an l_1 norm bound on the weight matrix results in a generalization error that is independent of the dimension of the data, 2) for any other norm constraint, there is a dimension dependence on the data.

For the first experiment, to show that imposing l_1 norm on weight matrix results in no dependence on the dimension of data, the following constrained empirical risk is minimized that essentially imposes a loose l_1 norm on the matrix.

$$\min_W \frac{1}{n}\sum_{i=1}^{n} \max_{x_i' \in B_x^\infty(\epsilon)} l(f_W(x_i'), y_i) + \lambda\|W\|_1, \tag{3}$$

where $l(\cdot)$ is the cross-entropy loss and $f_W(x) = Wx$.

Fig. 1 shows that under l_1 bounded norm, generalization error decreases as λ increase for any perturbation validating the implication of Thm. 1.

Fig. 2 shows that for $\lambda = 0$, generalization error increases as the dimension of data, d, increases, which also validates the result of Thm. 1.

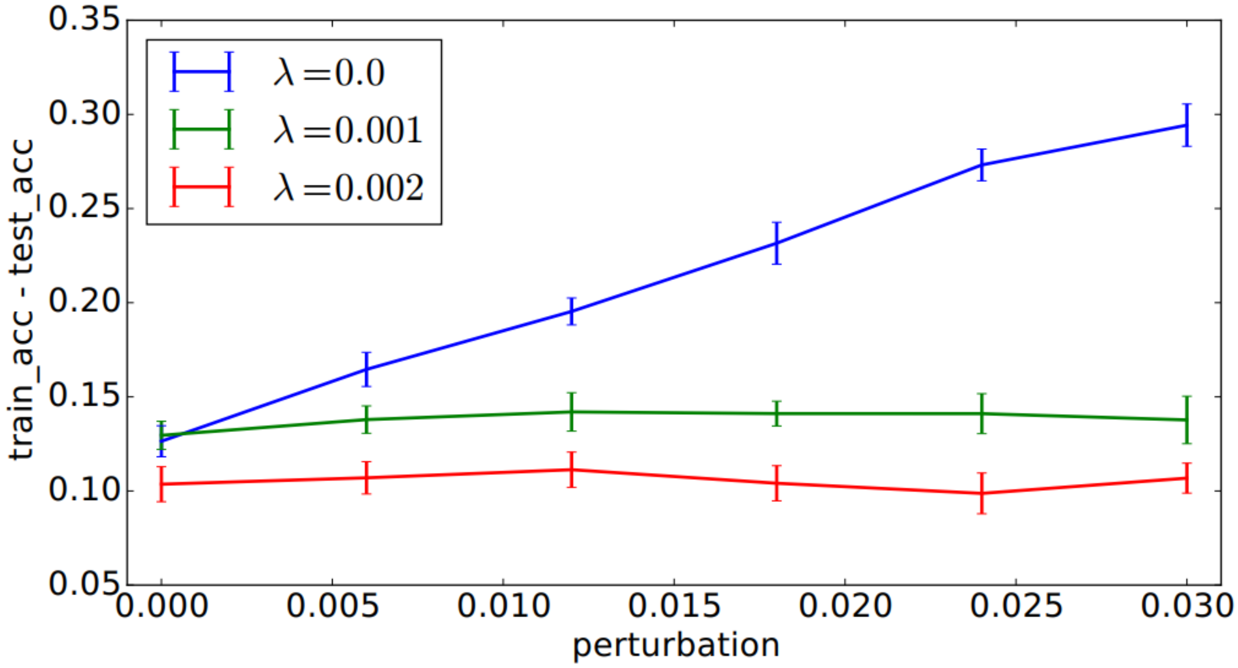

Figure 1: Generalization error decreases as λ increase for any given perturbation for linear classifiers [1].

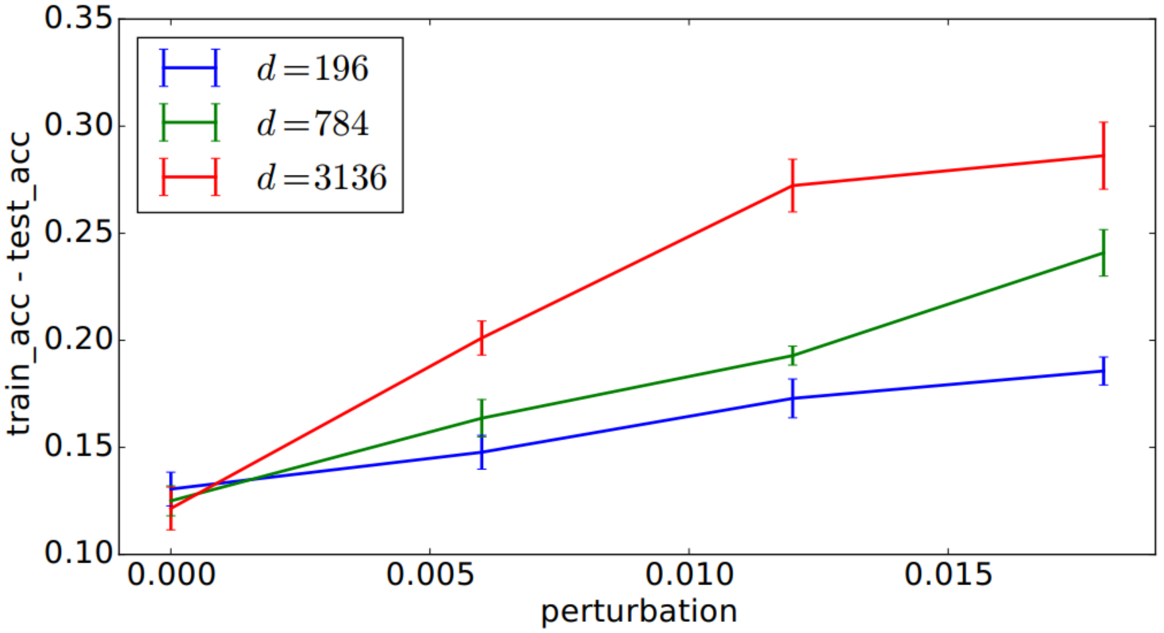

Figure 2: Generalization error increases as dimension of data, d, increases for any given perturbation for linear classifiers. Here $\lambda = 0$ [1].

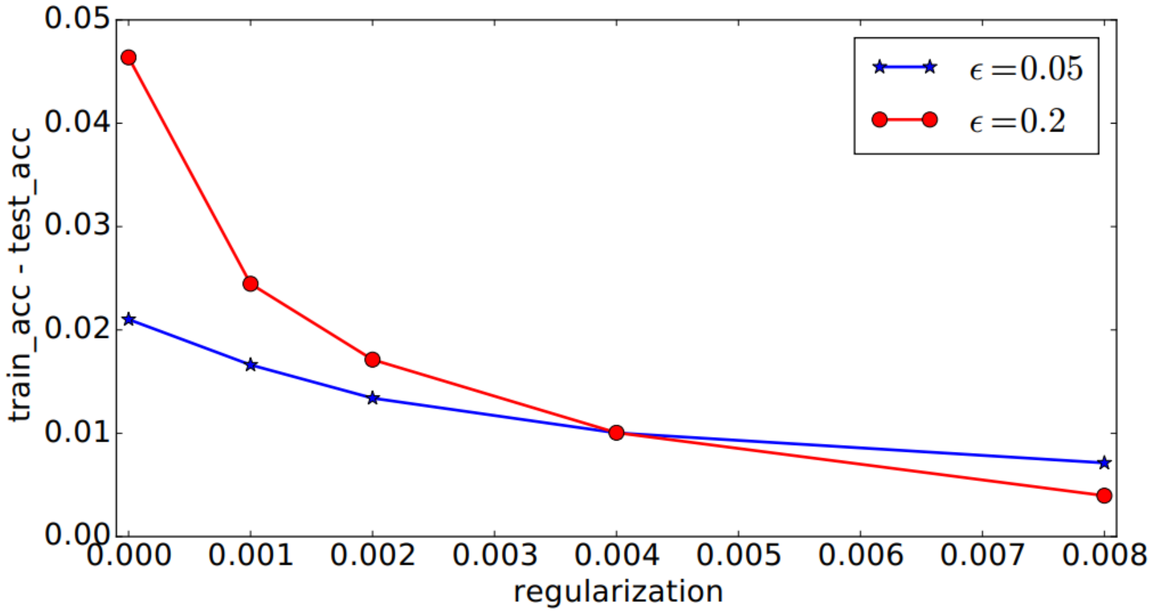

Figure 3: Adversarial generalization error vs regularization coefficient λ in neural networks.

5.2 Neural Networks

Now, an experiment is discussed that is performed on neural networks [1] to validate that l_1 norm can reduce the adversarial generalization error. The paper considers a four-layer ReLU neural network, where the first two layers are convolutional layers, whereas the last two layers are fully connected. The paper uses PGD attack [10] adversarial training to minimize the l_1 regularized objective function (3).

Fig. 3 shows the results of the experiment showing that as the regularization coefficient increases, the generalization error decreases for different values of perturbation, hence, verifying the claims in the paper.

6 Conclusion

The poor performance of state-of-the-art machine learning algorithms in the presence of adversaries has raised serious questions about their use in several critical applications such as self-driving cars or medical imaging. In the recent past, researchers have come up with several techniques to defend from particular type of adversarial attacks, however, it has been found that it is not very difficult to fool those defense algorithms. This calls for a theoretical treatment of the problem. Hence this report reviews a paper that finds bounds on the Rademacher complexity of linear classifiers as well as neural networks, taking a first step towards understanding the two main problems in adversarial learning: 1) optimizing the adversarial risk and 2) find the generalization error. The paper [1] shows that unless there is an l_1 norm bound on weight matrices for linear classifier, the Rademacher complexity depends on the dimension of the data, hence showing that l_1 norm might help in reducing generalization error under adversarial attacks. The paper also shows a similar dimension dependency when the classifiers are neural networks.

Acknowledgement

I would like to thank Prof. Bruce Hajeck for his guidance in choosing the topic and general discussions.

References

[1] Yin, Dong, Kannan Ramchandran, and Peter Bartlett. "Rademacher complexity for adversarially robust generalization." arXiv preprint arXiv:1810.11914 (2018).

[2] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[3] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013.

[4] Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." nature 529.7587 (2016): 484.

[5] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).

[6] Engstrom, Logan, et al. "A rotation and a translation suffice: Fooling cnns with simple transformations." arXiv preprint arXiv:1712.02779 (2017).

[7] Gilmer, Justin, et al. "Motivating the rules of the game for adversarial example research." arXiv preprint arXiv:1807.06732 (2018).

[8] Athalye, Anish, Nicholas Carlini, and David Wagner. "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples." arXiv preprint arXiv:1802.00420 (2018).

[9] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).dient policy for correlated normal beliefs." INFORMS journal on Computing 21.4 (2009): 599-613.

[10] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).dient policy for correlated normal beliefs." INFORMS journal on Computing 21.4 (2009): 599-613.

[11] Shaham, Uri, Yutaro Yamada, and Sahand Negahban. "Understanding adversarial training: Increasing local stability of supervised models through robust optimization." Neurocomputing 307 (2018): 195-204.

[12] Ledoux, Michel, and Michel Talagrand. Probability in Banach Spaces: isoperimetry and processes. Springer Science & Business Media, 2013.

[13] Cullina, Daniel, Arjun Nitin Bhagoji, and Prateek Mittal. "PAC-learning in the presence of evasion adversaries." arXiv preprint arXiv:1806.01471 (2018).

[14] Bartlett, Peter L., Dylan J. Foster, and Matus J. Telgarsky. "Spectrally-normalized margin bounds for neural networks." Advances in Neural Information Processing Systems. 2017.

[15] Raghunathan, Aditi, Jacob Steinhardt, and Percy Liang. "Certified defenses against adversarial examples." arXiv preprint arXiv:1801.09344 (2018).