Towards Understanding the Regularization of Adversarial Robustness on Neural Networks

Yuxin Wen [*1 2] **Shuai Li** [*1] **Kui Jia** [1 2]

Abstract

The problem of adversarial examples has shown that modern Neural Network (NN) models could be rather fragile. Among the more established techniques to solve the problem, one is to require the model to be ϵ-*adversarially robust* (AR); that is, to require the model not to change predicted labels when any given input examples are perturbed within a certain range. However, it is observed that such methods would lead to standard performance degradation, i.e., the degradation on natural examples. In this work, we study the degradation through the regularization perspective. We identify quantities from generalization analysis of NNs; with the identified quantities we empirically find that AR is achieved by regularizing/biasing NNs towards less confident solutions by making the changes in the feature space (induced by changes in the instance space) of most layers smoother uniformly in all directions; so to a certain extent, it prevents sudden change in prediction w.r.t. perturbations. However, the end result of such smoothing concentrates samples around decision boundaries, resulting in less confident solutions, and leads to worse standard performance. Our studies suggest that one might consider ways that build AR into NNs in a gentler way to avoid the problematic regularization.

1. Introduction

Despite the remarkable performance (Krizhevsky et al., 2012) of Deep Neural Networks (NNs), they are found to be rather fragile and easily fooled by adversarial examples (Szegedy et al., 2014). More intriguingly, these adversarial examples are generated by adding imperceptible noise to normal examples, and thus are indistinguishable for humans. NNs that are more robust to adversarial examples tend to have lower standard accuracy (Su et al., 2018), i.e., the accuracy measured on natural examples. The trade-off between robustness and accuracy has been empirically observed in many works (Fawzi et al., 2018; Kurakin et al., 2017; Madry et al., 2018; Tsipras et al., 2019), and has been theoretically analyzed under the context of simple models, e.g., linear models (Tsipras et al., 2019), quadratic models (Fawzi et al., 2018), but it is not clear whether the analysis generalizes to NNs. For example, Tsipras et al. (2019) show that for linear models, if examples are close to decision boundaries, robustness provably conflicts with accuracy, though the proof seems unlikely to generalize to NNs. Arguably, the most widely used remedy is developed to require NNs to be ϵ-*adversarially robust* (AR), e.g., via Adversarial Training (Madry et al., 2018), Lipschitz-Margin Training (Tsuzuku et al., 2018); that is, they require the model not to change predicted labels when any given input examples are perturbed within a certain range. In practice, such AR methods are found to lead to worse performance measured in standard classification accuracy. Alternatives to build AR into NNs are also being developed. For instance, Zhang et al. (2019) show that a gap exists between surrogate risk gap and 0-1 risk gap if many examples are close to decision boundaries, and better robustness can be achieved by pushing examples away from decision boundaries. But pushing examples away again degrades NN performance in their experiments. But they are yet to be widely adopted by the community.

We investigate how adversarial robustness built into NNs by the arguably most established method, i.e., Adversarial Training (Madry et al., 2018), influences the behaviors of NNs to make them more robust but have lower performance through the lens of regularization. In an earlier time (Szegedy et al., 2014), adversarial training has been suggested as a form of regularization: it augments the training of NNs with adversarial examples, and thus might improve the generalization of the end models. Note that such a *hard requirement* that the adversarial examples need to be classified correctly is different from the methods that increase

[*]Equal contribution [1]School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong 510640, China [2]Pazhou Lab, Guangzhou, 510335, China. Correspondence to: Shuai Li <lishuai918@gmail.com>, Yuxin Wen <wen.yuxin@mail.scut.edu.cn>, Kui Jia <kuijia@scut.edu.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

adversarial robustness by adding a soft penalty term to the risk function employed by Lyu et al. (2015) and Miyato et al. (2018), or a penalty term through curvature reduction (Moosavi-Dezfooli et al., 2019), or local linearization (Qin et al., 2019) (more discussion in appendix A). In these works, regularization is explicitly enforced by a penalty term, while in adversarial training, it is not clear that how training with augmented adversarial examples regularizes NNs. For example, if adversarial training does work as a regularizer, how does a possible improvement in generalization by using more data end up degrading performance? Even such a basic problem does not have a clear answer. To understand the regularization effects of AR on NNs, we go beyond simple linear or quadratic models and undertake a comprehensive generalization analysis of AR by establishing a rigorous generalization bound on NNs, and carrying out a series of empirical studies theoretically guided by the bound.

Technically, improved generalization implies the reduction in gap between training errors and test errors. Regularization achieves the gap reduction by reducing the size of the hypothesis space, which reduces the variance, but meanwhile increases the bias of prediction made — a constant classifier can have zero generalization errors, but also have low test performance. Thus, when a hypothesis space is improperly reduced, another possible outcome is biased poorly performing models with reduced generalization gaps.

Key results. Through a series of theoretically motivated experiments, we find that AR is achieved by regularizing/biasing NNs towards less confident solutions by making the changes in the feature space of most layers (which are induced by changes in the instance space) smoother uniformly in all directions; so to a certain extent, it prevents sudden change in prediction w.r.t. perturbations. However, the end result of such smoothing concentrates examples around decision boundaries and leads to worse standard performance. We elaborate the above statement in details shortly in section 1.1.

Implications. We conjecture that the improper reduction comes from the indistinguishability of the change induced in the intermediate layers of NNs by adversarial noise and that by inter-class difference. To guarantee AR, NNs are asked to smoothe out difference uniformly in all directions in a high dimensional space, and thus are biased towards less confident solutions that make similar/concentrated predictions. We leave the investigation of the conjecture as future works.

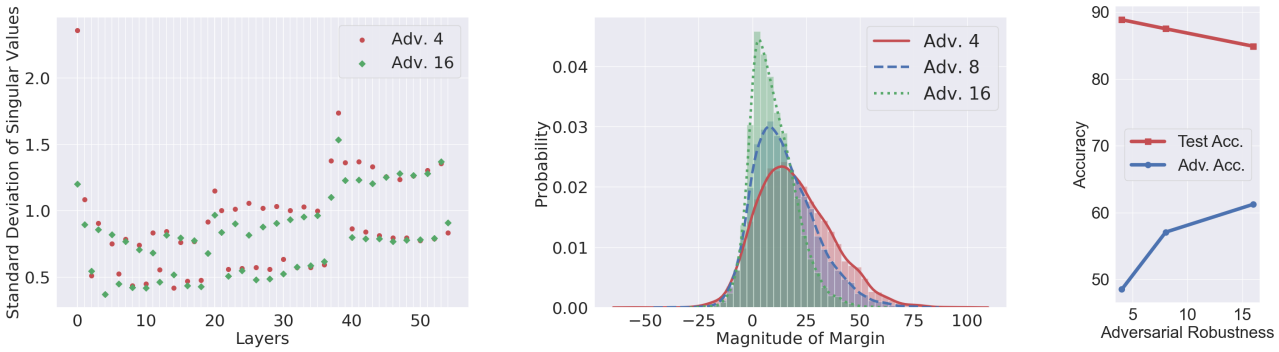
1.1. AR leads to less confident NNs with more indecisive misclassifications

This section elaborates the *key results* we briefly present previously.

AR reduces the perturbations in the activation/outputs — the perturbations that are induced by perturbations in the inputs fed into the layer — of most layers. Through a series of theoretically motivated experiments, the results prompt us to look at the singular value distributions of the weight matrix of each layer of the NNs. Shown in fig. 1a, we find that overall the standard deviation (STD) of singular values associated with a layer of the NN trained with lower AR strength 4 is larger than that of the NN with higher AR strength 16 [1] — the green dots are mostly below the red dots. Note that given a matrix $\boldsymbol{W}$ and an example $\boldsymbol{x}$, singular values of $\boldsymbol{W}$ determine how the norm $||\boldsymbol{W}\boldsymbol{x}||$ is changed comparing with $||\boldsymbol{x}||$. More specifically, let $\sigma_{\min}, \sigma_{\max}$ be the minimal and maximal singular values. If $\boldsymbol{x}$ is not in the null space of $\boldsymbol{W}$, then we have $||\boldsymbol{W}\boldsymbol{x}|| \in [\sigma_{\min}||\boldsymbol{x}||, \sigma_{\max}||\boldsymbol{x}||]$, where $||\cdot||$ denotes 2-norm. This applies to norm $||\delta\boldsymbol{x}||$ of a perturbation as well; that is, given possible changes $\delta\boldsymbol{x}$ of $\boldsymbol{x}$ of the same norm $||\delta\boldsymbol{x}|| = c$, where c is a constant, the variance of $\sigma(\boldsymbol{W})$ roughly determines the variance of $||\boldsymbol{W}\delta\boldsymbol{x}||$, where $\sigma(\boldsymbol{W})$ denotes all singular values $\{\sigma_i\}$ of $\boldsymbol{W}$. In more details, note that by SVD decomposition, $\boldsymbol{W}\delta\boldsymbol{x} = \sum_i \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \delta\boldsymbol{x}$, thus σ_i determines how the component $\boldsymbol{v}_i^T\delta\boldsymbol{x}$ in the direction of $\boldsymbol{v}_i$ is amplified. To see an example, suppose that $\sigma_{\min} = \sigma_{\max} = \sigma_0$, then the variance of $\sigma(\boldsymbol{W})$ is zero, and $||\boldsymbol{W}\delta\boldsymbol{x}|| = \sigma_0||\delta\boldsymbol{x}||$. In this case, the variance of $||\boldsymbol{W}\delta\boldsymbol{x}||$ (given an ensemble of perturbations $\delta\boldsymbol{x}$ of the same norm c) is zero as well. The conclusion holds as well for $\text{ReLU}(\boldsymbol{W}\delta x)$, where $\boldsymbol{W}$ here is a weight matrix of a layer of a NN, and ReLU denotes Rectifier Linear Unit activation function (proved by applying Cauchy interlacing law by row deletion (Chafai) in lemma 3.1). Consequently, by reducing the variance of singular values of weight matrix of a layer of the NN, AR reduces the variance of the norms of layer activations, or informally, perturbations in the activations, induced by input perturbations.

The perturbation reduction in activations concentrates examples, and it empirically concentrates them around decision boundaries; that is, predictions are less confident. The reduced variance implies that the outputs of each layer of the NN are more concentrated, but it does not tell where they are concentrated. Note that in the previous paragraph, the variance relationship discussed between $||\boldsymbol{W}\delta\boldsymbol{x}||$ and $||\delta\boldsymbol{x}||$ equally applies to $||\boldsymbol{W}\boldsymbol{x}||$ and $||\boldsymbol{x}||$, where $\boldsymbol{x}$ is an actual example instead of perturbations. Thus, to find out the concentration of perturbations, we can look at the concentration of samples. Technically, we look at *margins* of examples. In a multi-class setting, suppose a NN computes a score function $f : \mathbb{R}^d \to \mathbb{R}^L$, where L is the number of classes; a way

[1] The AR strength is characterized by the maximally allowed l_∞ norm of adversarial examples that are used to train the NNs — we use adversarial training (Madry et al., 2018) to build adversarial robustness into NNs. Details can be found in appendix B.1



(a) STD of Singular Values (b) Margin Distribution (c) Accuracy

Figure 1. Experiment results on ResNet56 (He et al., 2016) trained on the CIFAR10 dataset. For the details of the experiments, refer to section 4. **(a)** The standard deviation of singular values of each layer of NNs with adversarial robustness (AR) strength 4, 16 (AR strength 8 is dropped for clarity of the plot). To emphasize, the x-axis is the layer index — overall 56 layers are involved. **(b)** The probability distribution of margins of NNs with AR strength 4, 8, 16. **(c)** The standard and adversarial accuracy of NNs with AR 4, 8, 16.

to convert this to a classifier is to select the output coordinate with the largest magnitude, meaning $x \mapsto \arg\max_i f_i(x)$. The *confidence* of such a classifier could be quantified by margins. It measures the gap between the output for the correct label and other labels, meaning $f_y(x) - \max_{i \neq y} f_i(x)$. Margin piece-wise linearly depends on the scores, thus the variance of margins is also in a piece-wise linear relationship with the variance of the scores, which are computed linearly from the activation of a NN layer. Thus, the consequence of concentration of activation discussed in the previous paragraph can be observed in the distribution of margins. More details of the connection between singular values and margins are discussed in section 4.2.2, after we present lemma 3.1. A zero margin implies that a classifier has equal propensity to classify an example to two classes, and the example is on the decision boundary. We plot the margin distribution of the test set of CIFAR10 in fig. 1b, and find that margins are increasingly concentrated around zero — that is, the decision boundaries — as AR strength grows.

The sample concentration around decision boundaries smoothes sudden changes induced perturbations, but also increases indecisive misclassifications. The concentration of test set margins implies that the induced change in margins by the perturbation in the instance space is reduced by AR. Given two examples $\boldsymbol{x}, \boldsymbol{x}'$ from the test set, $\delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}'$ can be taken as a significant perturbation that changes the example $\boldsymbol{x}$ to $\boldsymbol{x}'$. The concentration of overall margins implies the change induced by $\delta\boldsymbol{x}$ is smaller statistically in NNs with higher AR strength. Thus, for an adversarial perturbation applied on $\boldsymbol{x}$, statistically the change of margins is smaller as well — experimentally it is reflected in the increased adversarial robustness of the network, as shown in the increasing curve in fig. 1c. That is, the sudden changes of margins originally induced by adversarial perturbations are *smoothed* (to change slowly). However, the *cost* of such smoothness is lower confidence in prediction, and more test examples are slightly/indecisively moved to the wrong sides of the decision boundaries — incurring lower accuracy, as shown in the decreasing curve in fig. 1c.

Lastly, we note that experiments in this section are used to illustrate our main arguments in this section. Further consistent quality results are reported in section 4 by conducting experiments on CIFAR10/100 and Tiny-ImageNet with networks of varied capacity. And more corroborative experiment results are presented in the appendices, and outlined in section 1.2.

1.2. Outline and contributions

This work carries out generalization analysis on NNs with AR. The quantities in the previous section are identified by the generalization errors (GE) upper bound we establish at theorem 3.1, which characterizes the regularization of AR on NNs. The key result is obtained at the *end* of a series of analysis, thus we present the outline of the analysis here.

Outline. After presenting some preliminaries in section 2, we proceed to analyze the regularization of AR on NNs, and establish a GE upper bound in section 3. The bound prompts us to look at the GE gaps in experiments. In section 4.1, we find that for NNs trained with higher AR strength, the surrogate risk gaps (GE gaps) decrease for a range of datasets, i.e., CIFAR10/100 and Tiny-ImageNet. It implies AR effectively regularizes NNs. We then study the finer behavior change of NNs that might lead to such a gap reduction. Again, we follow the guidance of theorem 3.1. We look at the margins in section 4.2.1, then at the singular value distribution in section 4.2.2, and discover the main results described in section 1.1. More corroborative experiments are present

in appendix B.4 and appendix B.6 to show that such phenomenon exists in a broad range of NNs with varied capacity and adversarial training techniques. More complementary results are present in appendix B.3 to explain some seemingly abnormal observations, and in appendix B.5 to quantitatively demonstrate the smoothing effects of AR discussed in section 1.1. Related works are present in appendix A.

Contributions. Overall, the core contribution in this work is to show that adversarial robustness (AR) regularizes NNs in a way that hurts its capacity to learn to perform in test. More specifically:

- We establish a generalization error (GE) bound that characterizes the regularization of AR on NNs. The bound connects *margin* with adversarial robustness radius ϵ via *singular values of weight matrices* of NNs, thus suggesting the two quantities that guide us to investigate the regularization effects of AR empirically.
- Our empirical analysis tells that AR *effectively* regularizes NNs to reduce the GE gaps. To understand how reduced GE gaps turns out to degrade test performance, we study *variance of singular values* of layer-wise weight matrices of NNs and *distributions of margins* of samples, when different strength of AR are applied on NNs.
- The study shows that AR is achieved by regularizing/biasing NNs towards less confident solutions by making the changes in the feature space of most layers (which are induced by changes in the instance space) smoother uniformly in all directions; so to a certain extent, it prevents sudden change in prediction w.r.t. perturbations. However, the end result of such smoothing concentrates samples around decision boundaries and leads to worse standard performance.

2. Preliminaries

Assume an instance space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the space of input data, and $\mathcal{Y}$ is the label space. $Z := (X, Y)$ are the random variables with an unknown distribution μ, from which we draw samples. We use $S_m = \{z_i = (\boldsymbol{x}_i, y_i)\}_{i=1}^m$ to denote the training set of size m whose examples are drawn independently and identically distributed (i.i.d.) by sampling Z. Given a loss function l, the goal of learning is to identify a function $T : \mathcal{X} \mapsto \mathcal{Y}$ in a hypothesis space (a class $\mathcal{T}$ of functions) that minimizes the expected risk

$$R(l \circ T) = \mathbb{E}_{Z\sim\mu}\left[l\left(T(X), Y\right)\right],$$

Since μ is unknown, the observable quantity serving as the proxy to the expected risk R is the empirical risk

$$R_m(l \circ T) = \frac{1}{m}\sum_{i=1}^{m} l\left(T(\boldsymbol{x}_i), y_i\right).$$

Our goal is to study the discrepancy between R and R_m, which is termed as *generalization error* — it is also sometimes termed as generalization gap in the literature

$$\mathrm{GE}(l \circ T) = |R(l \circ T) - R_m(l \circ T)|. \tag{1}$$

A NN is a map that takes an input x from the space $\mathcal{X}$, and builds its output by recursively applying a linear map W_i followed by a pointwise non-linearity g:

$$x_i = g(\boldsymbol{W}_i \boldsymbol{x}_{i-1}),$$

where i indexes the times of recursion, which is denoted as a layer in the community, $i = 1, \ldots, L$, $x_0 = x$, and g denotes the activation function. which is restricted to Rectifier Linear Unit (ReLU) (Glorot et al., 2011) or max pooling operator (Bécigneul, 2017) in this paper. To compactly summarize the operation of T, we denote

$$Tx = g(\boldsymbol{W}_L g(\boldsymbol{W}_{L-1} \ldots g(\boldsymbol{W}_1 \boldsymbol{x}))). \tag{2}$$

Definition 1 (Covering number)**.** *Given a metric space $(\mathcal{S}, \rho)$, and a subset $\tilde{\mathcal{S}} \subset \mathcal{S}$, we say that a subset $\hat{\mathcal{S}}$ of $\tilde{\mathcal{S}}$ is a ϵ-cover of $\tilde{\mathcal{S}}$, if $\forall \tilde{s} \in \tilde{\mathcal{S}}$, $\exists \hat{s} \in \hat{\mathcal{S}}$ such that $\rho(\tilde{s}, \hat{s}) \leq \epsilon$. The ϵ-covering number of $\tilde{\mathcal{S}}$ is*

$$\mathcal{N}_\epsilon(\tilde{\mathcal{S}}, \rho) = \min\{|\hat{\mathcal{S}}| : \hat{\mathcal{S}} \text{ is an } \epsilon\text{-covering of } \tilde{\mathcal{S}}\}.$$

Various notions of adversarial robustness have been studied in existing works (Madry et al., 2018; Tsipras et al., 2019; Zhang et al., 2019). They are conceptually similar; in this work, we formalize its definition to make clear the object for study.

Definition 2 ((ρ, ϵ)-adversarial robustness)**.** *Given a multi-class classifier $f : \mathcal{X} \to \mathbb{R}^L$, and a metric ρ on $\mathcal{X}$, where L is the number of classes, f is said to be adversarially robust w.r.t. adversarial perturbation of strength ϵ, if there exists an $\epsilon > 0$ such that $\forall z = (\boldsymbol{x}, y) \in \mathcal{Z}$ and $\delta\boldsymbol{x} \in \{\rho(\delta\boldsymbol{x}) \leq \epsilon\}$, we have*

$$f_{\hat{y}}(\boldsymbol{x} + \delta\boldsymbol{x}) - f_i(\boldsymbol{x} + \delta\boldsymbol{x}) \geq 0,$$

*where $\hat{y} = \arg\max_j f_j(\boldsymbol{x})$ and $i \neq \hat{y} \in \mathcal{Y}$. ϵ is called **adversarial robustness radius**. When the metric used is clear, we also refer (ρ, ϵ)-adversarial robustness as ϵ-adversarial robustness.*

Note that the definition is an *example-wise* one; that is, it requires each example to have a guarding area, in which all examples are of the same class. Also note that the robustness is w.r.t. the predicted class, since ground-truth label is unknown for a $\boldsymbol{x}$ in test.

We characterize the GE with ramp risk, which is a typical risk to undertake theoretical analysis (Bartlett et al., 2017; Neyshabur et al., 2018b).

Definition 3 (Margin Operator)**.** *A margin operator $\mathcal{M} : \mathbb{R}^L \times \{1, \ldots, L\} \to \mathbb{R}$ is defined as*

$$\mathcal{M}(\boldsymbol{s}, y) := s_y - \max_{i \neq y} s_i$$

Definition 4 (Ramp Loss). *The ramp loss $l_\gamma : \mathbb{R} \to \mathbb{R}^+$ is defined as*

$$l_\gamma(r) := \begin{cases} 0 & r < -\gamma \\ 1 + r/\gamma & r \in [-\gamma, 0] \\ 1 & r > 0 \end{cases}$$

Definition 5 (Ramp Risk). *Given a classifier f, ramp risk is the risk defined as*

$$R_\gamma(f) := \mathbb{E}(l_\gamma(-\mathcal{M}(f(X), Y))),$$

where X, Y are random variables in the instance space $\mathcal{Z}$ previously.

We will use a different notion of margin in theorem 3.1, and formalize its definition as follows. We reserve the unqualified word "margin" specifically for the margin discussed previously — the output of margin operator for classification. We call this margin to-be-introduced *instance-space margin (IM)*.

Definition 6 (Smallest Instance-space Margin). *Given an element $z = (\boldsymbol{x}, y) \in \mathcal{Z}$, let $v(\boldsymbol{x})$ be the distance from $\boldsymbol{x}$ to its closest point on the decision boundary, i.e., the instance-space margin (IM) of example $\boldsymbol{x}$. Given a covering set $\hat{S}$ of $\mathcal{Z}$, let*

$$v_{\min} = \min_{\boldsymbol{x} \in \{\boldsymbol{x} \in \mathcal{X} | \exists \boldsymbol{x}' \in \hat{S}_m, ||\boldsymbol{x} - \boldsymbol{x}'||_2 \le \epsilon\}} v(\boldsymbol{x}), \tag{3}$$

where $\hat{S}_m := \{\boldsymbol{x}' \in \hat{S} | \exists \boldsymbol{x}_i \in S_m, ||\boldsymbol{x}_i - \boldsymbol{x}'||_2 \le \epsilon\}$. $v_{\min}$ is the smallest instance-space margin of elements in the covering balls that contain training examples.

3. Theoretical instruments for empirical studies on AR

In this section, we rigorously establish the bound mentioned in the introduction. We study the map T defined in section 2 as a NN (though technically, T now is a map from $\mathcal{X}$ to $\mathbb{R}^L$, instead of to $\mathcal{Y}$, such an abuse of notation should be clear in the context). To begin with, we introduce an assumption, before we state the generalization error bound guaranteed by adversarial robustness.

Assumption 3.1 (Monotony). *Given a point $\boldsymbol{x} \in \mathcal{X}$, let $\boldsymbol{x}'$ be the point on the decision boundary of a NN T that is closest to $\boldsymbol{x}$. Then, for all $\boldsymbol{x}''$ on the line segment $\boldsymbol{x} + t(\boldsymbol{x}' - \boldsymbol{x}), t \in [0, 1]$, the margin $\mathcal{M}(T\boldsymbol{x}'', y)$ decreases monotonously.*

The assumption is a regularity condition on the classifier that rules out undesired oscillation between $\boldsymbol{x}$ and $\boldsymbol{x}'$. To see how, notice that the margin defined in definition 3 reflects how confident the decision is made. Since $\boldsymbol{x}'$ is on the decision boundary, it means the classifier is unsure how it should be classified. Thus, when the difference $\boldsymbol{x}' - \boldsymbol{x}$ is gradually added to $\boldsymbol{x}$, ideally we want the confidence that we have on classifying $\boldsymbol{x}$ to decrease in a consistent way to reflect the uncertainty.

Theorem 3.1. *Let T denote a NN with ReLU and MaxPooling nonlinear activation functions (the definition is put at eq. (2) for readers' convenience), l_γ the ramp loss defined at definition 4, and $\mathcal{Z}$ the instance space assumed in section 3. Assume that $\mathcal{Z}$ is a k-dimensional regular manifold that accepts an ϵ-covering with covering number $(\frac{C_{\mathcal{X}}}{\epsilon})^k$, and assumption assumption 3.1 holds. If T is ϵ_0-adversarially robust (defined at definition 2), $\epsilon \le \epsilon_0$, and denote $v_{\min}$ the smallest IM margin in the covering balls that contain training examples (defined at definition 6), $\sigma^i_{\min}$ the smallest singular values of weight matrices $\boldsymbol{W}_i, i = 1, \ldots, L-1$ of a NN, $\{\boldsymbol{w}_i\}_{i=1,\ldots,|\mathcal{Y}|}$ the set of vectors made up with ith rows of $\boldsymbol{W}_L$ (the last layer's weight matrix), then given an i.i.d. training sample $S_m = \{z_i = (\boldsymbol{x}_i, y_i)\}_{i=1}^m$ drawn from $\mathcal{Z}$, its generalization error GE$(l \circ T)$ (defined at eq. (1)) satisfies that, for any $\eta > 0$, with probability at least $1 - \eta$*

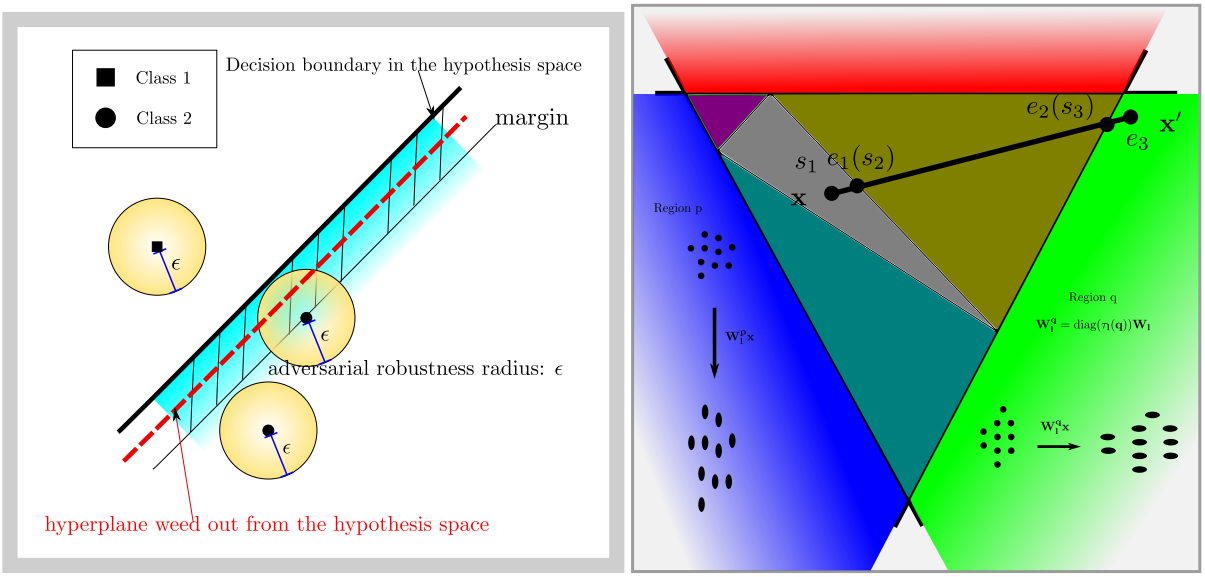
$$\begin{aligned} GE(l_\gamma \circ T) \le & \max\{0, 1 - \frac{u_{\min}}{\gamma}\} \\ & + \sqrt{\frac{2\log(2)C^k_{\mathcal{X}}}{\epsilon^k m} + \frac{2\log(1/\eta)}{m}} \end{aligned} \tag{4}$$

where

$$u_{\min} = \min_{y, \hat{y} \in \mathcal{Y}, y \ne \hat{y}} ||\boldsymbol{w}_y - \boldsymbol{w}_{\hat{y}}||_2 \prod_{i=1}^{L-1} \sigma^i_{\min} v_{\min} \tag{5}$$

is a lower bound of margins of examples in covering balls that contain training samples.

The proof of theorem 3.1 is in appendix C. *The bound identifies quantities that would be studied experimentally in section 4 to understand the regularization of AR on NNs.* The first term in eq. (4) in theorem 3.1 suggests that quantities related to the lower bound of *margin* $u_{\min}$ might be useful to study how ϵ-adversarial robustness (ϵ-AR) regularizes NNs. However, ϵ-AR is guaranteed in the instance space that determines the smallest *instance-space margin* $v_{\min}$. To relate GE bound with ϵ-AR, we characterize in eq. (5) the relationship between margin with IM, via smallest *singular values of NNs' weight matrices*, suggesting that quantities related to singular values of NNs' weight matrices might be useful to study how AR regularizes NNs as well. An illustration on how AR could influence generalization of NNs through IM is also given in fig. 2a. The rightmost term in eq. (4) is a standard term in robust framework (Xu & Mannor, 2012) in learning theory, and is not very relevant to the discussion. *The remaining of this paper are empirical studies that are based on the quantities*, e.g., margin distributions and singular values of NNs' weight matrices, that are related to the identified quantities, i.e., $u_{\min}, \sigma^i_{\min}$.



(a) (b)

Figure 2. **(a)** Illustration of the regularization effect of adversarial robustness. If a NN T is ϵ-adversarially robust, for a given example $\boldsymbol{x}$ (drawn as filled squares or circles) and points $\boldsymbol{x}'$ in the yellow ball $\{\boldsymbol{x}' \mid \rho(\boldsymbol{x}, \boldsymbol{x}') \leq \epsilon\}$ around $\boldsymbol{x}$, the predicted labels of $\boldsymbol{x}, \boldsymbol{x}'$ should be the same, and the loss variation is potentially bigger as $\boldsymbol{x}'$ moves from the center to the edge, as shown as intenser yellow color at the edge of a ball. Collectively, the adversarial robustness of each example requires an *instance-space margin* (IM) to exist for the decision boundary, shown as the shaded cyan margin. As normally known, margin is related to generalization ability that shrinks the hypothesis space. In this case, the IM required by adversarial robustness would weed out hypotheses that do not have an adequate IM, such as the red dashed line shown in the illustration. **(b)** Illustration of lemma 3.1. Given a NN with ReLU activation function, the feature map $\boldsymbol{I}_l$ at layer l is divided into regions where $\boldsymbol{I}_l(\boldsymbol{x})$ is piecewise linear w.r.t. $\boldsymbol{x}$. The induced linear map $\boldsymbol{W}_1^q$ is given by $\text{diag}(\tau_1(q))\boldsymbol{W}_1$, where $\text{diag}(\tau_l(q))$ is a diagonal matrix whose diagonal entries are given by a vector $\tau_l(q)$ that has 0-1 values. For example, in region p, $\boldsymbol{I}_1 = \boldsymbol{W}_1^p\boldsymbol{x}$ and distance between instances $\boldsymbol{x}$ are vertical elongated, while in region q, $\boldsymbol{I}_1 = \boldsymbol{W}_1^q\boldsymbol{x}$ and distance are horizontally elongated. Thus given $\boldsymbol{x}, \boldsymbol{x}'$, the difference $||\boldsymbol{I}_l(\boldsymbol{x}) - \boldsymbol{I}_l(\boldsymbol{x}')||$ between $\boldsymbol{I}_l(\boldsymbol{x})$ and $\boldsymbol{I}_l(\boldsymbol{x}')$ is the length of the transformed line segment $\boldsymbol{x} - \boldsymbol{x}'$ drawn, of which each segment is linearly transformed in a different way.

These studies aim to illuminate with empirical evidence on the phenomena that AR regularizes NNs, reduces GE gaps, but degrades test performance. [2]

Before turning into empirical study, we further present a lemma to illustrate the relation characterized in eq. (5) without the need to jump into proof of theorem 3.1. It would motivate our experiments later in section 4.2.2. We state the following lemma that relates distances between elements in the instance space with those in the feature space of any intermediate network layers.

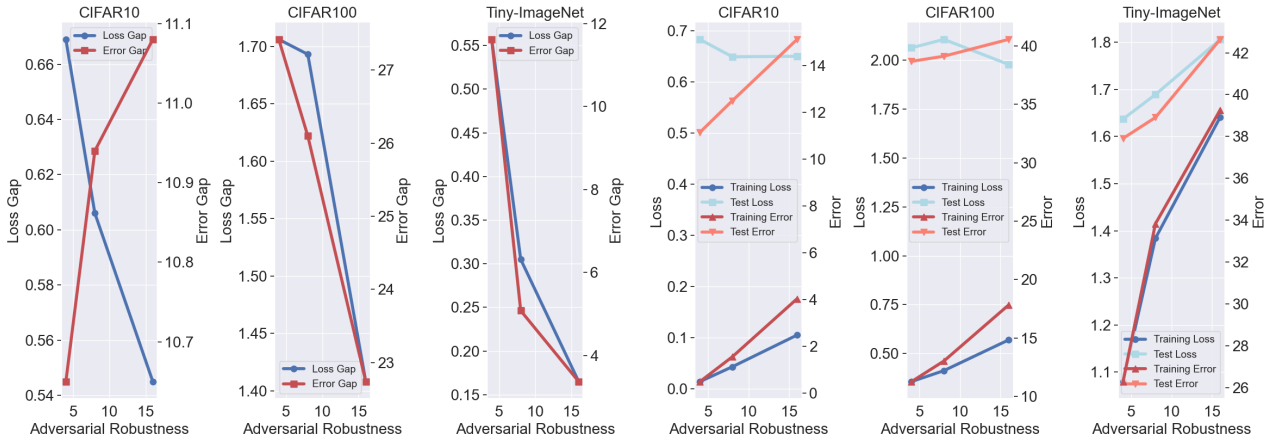
Lemma 3.1. *Given two instances $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, let $\boldsymbol{I}_l(\boldsymbol{x})$ be the activation $g(\boldsymbol{W}_l g(\boldsymbol{W}_{l-1} \dots g(\boldsymbol{W}_1\boldsymbol{x})))$ at layer l of $\boldsymbol{x}$, then there exist $n \in \mathbb{N}$ sets of matrices $\{\boldsymbol{W}_i^{q_j}\}_{i=1\dots l}, j = 1 \dots n$, that each of the matrix $\boldsymbol{W}_i^{q_j}$ is obtained by setting some rows of $\boldsymbol{W}_i$ to zero, and $\{q_j\}_{j=1\dots n}$ are arbitrary distinctive symbols indexed by j that index $\boldsymbol{W}_i^{q_j}$, such that*

$$||\boldsymbol{I}_l(\boldsymbol{x}) - \boldsymbol{I}_l(\boldsymbol{x}')|| = \sum_{j=1}^{n} \int_{s_j}^{e_j} \prod_{i=1}^{l} \boldsymbol{W}_i^{q_j} dt(\boldsymbol{x} - \boldsymbol{x}')$$

where $s_1 = 0, s_{j+1} = e_j, e_n = 1, s_j, e_j \in [0, 1]$ — each $[s_j, e_j]$ is a segment in the line segment parameterized by t that connects $\boldsymbol{x}$ and $\boldsymbol{x}'$.

Its proof is in appendix C, and an illustration is given in fig. 2b. Essentially, it states that difference in the feature space of a NN, induced by the difference between elements in the instance space, is a summation of the norms of the linear transformation ($\prod_{i=1}^{l} \boldsymbol{W}_i^{q_j}$) applied on segments of the line segment that connects $\boldsymbol{x}, \boldsymbol{x}'$ in the instance space.

[2] Note that in the previous paragraph, though we identifies quantities $u_{\min}$ and $\sigma_{\min}^i$ related to the upper bound of GE, the quantities we actually would study empirically are *margin distribution* and all *singular values* that characterize the GE of all samples, not just the extreme case (upper bound). The analytic characterization of the GE of all samples is not possible since we do not have enough information (we do not know the true distribution of samples). That's why to arrive at close-form analytic characterization of GE, we resort to the extreme non-asymptotic large-sample behaviors. *The analytic form is a neat way to present how relevant quantities influence GE.* In the rest of the paper, we would carry on empirical study on the distributions of margins and singular values to investigate AR's influence on GE of all samples.



(a) (b)

Figure 3. Experiment results on CIFAR10/100, and Tiny-ImageNet. The unit of x-axis is the adversarial robustness (AR) strength of NNs, c.f. the beginning of section 4. **(a)** Plots of loss gap (and error rate gap) between training and test datasets v.s. AR strength. **(b)** Plots of losses (and error rates) on training and test datasets v.s. AR strength.

Since $\boldsymbol{W}_i^{q_j}$ is obtained by setting rows of $\boldsymbol{W}_i$ to zero, the singular values of these induced matrices are intimately related to weight matrices $\boldsymbol{W}_i$ of NN by Cauchy interlacing law by row deletion (Chafai). Since the margin of an example $\boldsymbol{x}$ is a linear transform of the difference between $I_{L-1}(\boldsymbol{x})$ and the $I_{L-1}(\boldsymbol{x}')$ of an element $\boldsymbol{x}'$ on the decision boundary, singular values of $\{W_i\}_{i=1\ldots L-1}$ determine the amplification/shrinkage of the IM $\boldsymbol{x} - \boldsymbol{x}'$.

4. Empirical studies on regularization of adversarial robustness

In this section, guided by theorem 3.1, we undertake empirical studies to explore AR's regularization effects on NNs. We first investigate the behaviors of off-the-shelf architectures of fixed capacity on various datasets in section 4.1 and section 4.2. More corroborative controlled studies that explore the regularization effects of AR on NNs with varied capacity are present in appendix B.3.

4.1. Adversarial robustness effectively regularizes NNs on various datasets

This section aims to explore whether AR can effectively reduce generalization errors — more specifically, the surrogate risk gaps. We use adversarial training (Madry et al., 2018) to build adversarial robustness into NNs. The AR strength is characterized by the maximally allowed l_∞ norm of adversarial examples that are used to train the NNs. Details on the technique to build adversarial robustness into NNs is given in appendix B.1.

Our experiments are conducted on CIFAR10, CIFAR100, and Tiny-ImageNet (ImageNet, 2018) that represent learning tasks of increased difficulties. We use ResNet-56 and ResNet-110 (He et al., 2016) for CIFAR10/100, and Wide ResNet (WRN-50-2-bottleneck) (Zagoruyko & Komodakis, 2016) for Tiny-ImageNet (ImageNet, 2018). These networks are trained with increasing AR strength. Results are plotted in fig. 3.

Regularization of AR on NNs. We observe in fig. 3a (shown as blue lines marked by circles) that GE gaps (the gaps between training and test losses) decrease as strength of AR increase; we also observe in fig. 3a that training losses increase as AR strength increase; these results (and more results in subsequent fig. 6) imply that AR does regularize training of NNs by reducing their capacities to fit training samples. Interestingly, in the CIFAR10/100 results in fig. 3b, the test losses show a decreasing trend even when test error rates increase. It suggests that the network actually performs better measured in test loss as contrast to the performance measured in test error rates. This phenomenon results from that less confident wrong predictions are made by NNs thanks to adversarial training, which will be explained in details in section 4.2, when we carry on finer analysis. We note that on Tiny-ImageNet, the test loss does not decrease as those on CIFAR10/100. It is likely because the task is considerably harder, and regularization hurts NNs even measured in test loss.

Trade-off between regularization of AR and test error rates. The error rate curves in fig. 3b also tell that the end result of AR regularization leads to biased-performing NNs that achieve degraded test performance. These results are consistent across datasets and networks.

Seemingly abnormal phenomenon. An seemingly abnormal phenomenon in CIFAR10 observed in fig. 3a is that the error rate gap actually increases. It results from the same underlying behaviors of NNs, which we would introduce in section 4.2, and an overfitting phenomenon that AR cannot control. Since it would be a digress to explain, it is put in appendix B.3.

We finally note that the adversarial robustness training reproduced is relevant, of which the defense effect is comparable with existing works. One may refer to fig. 12 in appendix D.2 for the details. We can see from it that similar adversarial robustness to Madry et al. (2018) and Li et al. (2018) is achieved for CIFAR10/100, Tiny-ImageNet in the NNs we reproduce.

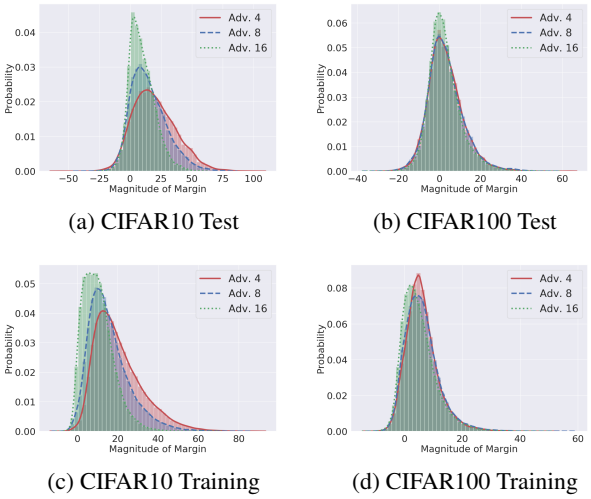
4.2. Refined analysis through margins and singular values

The experiments in the previous sections confirm that AR reduces GE, but decreases accuracy. We study the underlying behaviors of NNs to analyze what have led to it here. More specifically, we show that adversarial training implements ϵ-adversarial robustness by making NNs biased towards less confident solutions; that is, the key finding we present in section 1.1 that explains both the prevented sudden change in prediction w.r.t. sample perturbation (i.e., the achieved AR), and the reduced test accuracy.

4.2.1. Margins that concentrate more around zero lead to reduced GE gap

To study how GE gaps are reduced, theorem 3.1 suggests we first look at the margins of examples — a lower bound of margins is $u_{\min}$ in eq. (5). The analysis on margins has been a widely used tool in learning theory (Bartlett et al., 2017). It reflects the confidence that a classifier has on an example, which after being transformed by a loss function, is the surrogate loss. Thus, the loss difference between examples are intuitively reflected in the difference in confidence characterized by margins. To study how AR influences generalization of NNs, distributions of samples which are obtained by training ResNet-56 on CIFAR10 and CIFAR100 with increased AR strength (the same setting as for fig. 3). Applying the same network of ResNet-56 respectively on *CIFAR-10 and CIFAR-100 of different learning difficulties* creates learning settings of *larger- and smaller-capacity* NNs.

Concentration and reduced accuracy. In fig. 4, we can see that in both CIFAR10/100, the distributions of margins become more concentrated around zero as AR grows. The concentration moves the mode of margin distribution towards zero and more examples slightly across the decision boundaries, where the margins are zero, which explains the reduced accuracy [3].



(a) CIFAR10 Test (b) CIFAR100 Test

(c) CIFAR10 Training (d) CIFAR100 Training

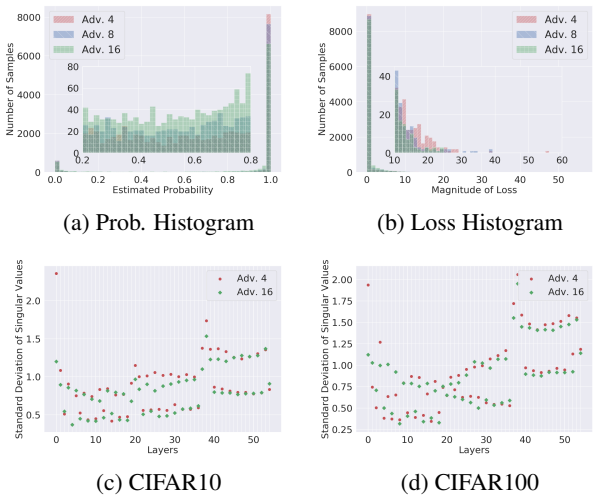
Figure 4. Margin distributions of NNs with AR strength 4, 8, 16 on Training and Test sets of CIFAR10/100.

Concentration and reduced loss/GE gap. The concentration has different consequences on training and test losses. Before describing the consequences, to directly relate the concentration to loss gap, we further introduce *estimated probabilities* of examples. This is because though we use ramp loss in theoretical analysis, in the experiments, we explore the behaviors of more practically used *cross entropy loss*. The loss maps one-to-one to estimated probability, but not to margin, though they both serve as a measure of confidence. Suppose $\boldsymbol{p}(\boldsymbol{x})$ is the output of the *softmax* function of dimension L (L is the number of target classes), and y is the target label. The estimated probability of $\boldsymbol{x}$ would be the y-th dimension of $(\boldsymbol{p}(\boldsymbol{x}))$, i.e., $(\boldsymbol{p}(\boldsymbol{x}))_y$. **On the training sets**, since the NNs are optimized to perform well on the sets, only a tiny fraction of them are classified wrongly. To concentrate the margin distribution more around zero, is to

[3] We remark a possibly confusing phenomenon here about the margin. The bound eq. (4) might give the impression that a smaller margin might lead to a larger generalization error, while the empirical study instead shows that the NNs with a smaller margin have a smaller generalization error. The hypothesized confusion is a misunderstanding of the generalization bound analysis. The upper bound is a worst case analysis of GE. However, in practice, the interesting object is the average gap between the training losses and the test losses, i.e., the GE. Unfortunately, the average gap cannot be analyzed analytically (cf. footnote 2). Thus, we, and also the statistical learning community, resort to worst case analysis to find an upper bound on GE to identify quantities that might influence GE. In this case, the phenomenon suggests that the bound might be loose, though this is a problem that plagues the statistical learning community (Nagarajan & Kolter, 2019). But our focus in this work is not to derive tight bounds, or reach definite conclusions from bounds alone, but to guide experiments with the bound.

make almost all of predictions that are correct less confident. Thus, a higher expected training loss ensues. **On the test sets**, the estimated probabilities of the target class concentrate more around middle values, resulting from lower confidence/margins in predictions made by NNs, as shown in fig. 5a (but the majority of values are still at the ends). Note that wrong predictions away from decision boundaries (with large negative margins) map to large loss values in the surrogate loss function. Thus, though NNs with larger AR strength have lower accuracy, they give more predictions whose estimated probabilities are at the middle (compared with NNs with smaller AR strength). These predictions, even if relatively more of them are wrong, maps to smaller loss values, as shown in fig. 5b, where we plot the histogram of loss values of test samples. In the end, it results in expected test losses that are lower, or increase in a lower rate than the training losses on CIFAR10/100, Tiny-ImageNet, as shown in fig. 3b. **The reduced GE gap** results from the increased training losses, and decreased or less increased test losses.

4.2.2. AR MAKES NNS SMOOTHE PREDICTIONS W.R.T. INPUT PERTURABTIONS IN ALL DIRECTIONS



(a) Prob. Histogram (b) Loss Histogram

(c) CIFAR10 (d) CIFAR100

Figure 5. **(a)(b)** are histograms of estimated probabilities and losses of the test set sample of NNs trained with AR strength 4, 8, 16. We plot a subplot of a narrower range inside the plot of the full range to better show the histograms of examples that are around the middle values induced by AR. **(c)(d)** are standard deviations of singular values of weight matrices of NNs at each layer trained on CIFAR10/100 with AR strength 4, 16. The AR strength 8 is dropped for clarity.

The observation in section 4.2.1 shows that AR make NNs just *less confident* by reducing the variance of predictions made and concentrate margins more around zero. In this section, we study the *underlying factors* of AR that make NNs become less confident.

To begin with, we show that the singular values of the weight matrix of each layer determine the perturbation in margins of samples induced by perturbations in the instance space. Such a connection between singular values and the perturbation of outputs of a single layer, i.e., $\text{ReLU}(\boldsymbol{W}\delta\boldsymbol{x})$, has been discussed in section 1.1. In the following, with lemma 3.1, we describe how the relatively more complex connection between margins and singular values of each weight matrix of layers of NNs holds. Observe that margins are obtained by applying a piece-wise linear mapping (c.f. the margin operator in definition 3) to the activation of the last layer of a NN. It implies the perturbations in activation of the last layer induce changes in margins in a piece-wise linear way. Meanwhile, the perturbation in the activation of the last layer (induced by perturbation in the instance space) is determined by the weight matrix's singular values of each layer of NNs. More specifically, this is explained as follows. Lemma 3.1 shows that the perturbation $\delta\boldsymbol{I}$ induced by $\delta\boldsymbol{x}$, is given by $\sum_{j=1}^{n}\int_{s_j}^{e_j}\prod_{i=1}^{l}\boldsymbol{W}_i^{q_j}\delta\boldsymbol{x}dt$. Note that for each i, $\boldsymbol{W}_i^{q_i}$ is a matrix. By Cauchy interlacing law by row deletion (Chafai), the singular values of $\boldsymbol{W}_i$, the weight matrix of layer i, determine the singular values of $\boldsymbol{W}_i^{q_j}$. Thus, suppose $l = 1$, we have the change (measured in norm) induced by perturbation as $\sum_{j=1}^{n}\int_{s_j}^{e_j}\boldsymbol{W}_1^{q_j}\delta\boldsymbol{x}dt$. The singular values of $\boldsymbol{W}_1$ would determine the variance (of norms) of activation perturbations induced by perturbations $\delta\boldsymbol{x}$, similarly as explained in section 1.1 except that the norm perturbation now is obtained by a summation of n terms $\boldsymbol{W}_1^{q_j}\delta\boldsymbol{x}dt$ (each of which is the exact form discussed in section 1.1) weighted by $1/(e_j - s_j)$. Similarly, for the case where $l = 2 \ldots L-1$, the singular values of $\boldsymbol{W}_l$ determine the variance of perturbations in the output of layer l that induced by the perturbations in the output of the previous layer (the input to layer l), i.e., layer $l-1$. Consequently, we choose to study these singular values.

We show the standard deviation of singular values of each layer of ResNet56 trained on CIFAR10/100 earlier in fig. 5c fig. 5d. Overall, the standard deviation of singular values associated with a layer of the NN trained with AR strength 4 is mostly larger than that of the NN with AR strength 16. The STD reduction in CIFAR100 is relatively smaller than CIFAR10, since as observed in fig. 4b, the AR induced concentration effect of margin distributions is also relatively less obvious than that in fig. 4a. More quantitative analysis is given in appendix B.2. This leads us to our *key results* described in section 1.1.

Acknowledgements

This work is supported in part by National Natural Science Foundation of China (Grant No. 61771201), Program for Guangdong Introducing Innovative and Enterpreneurial Teams (Grant No. 2017ZT07X183), Guangdong R&D

Key Project of China (Grant No. 2019B010155001) and Guangzhou Key Laboratory of Body Data Science (Grant No. 201605030011).

References

Attias, I., Kontorovich, A., and Mansour, Y. Improved generalization bounds for robust learning. Technical report, 2018. URL https://arxiv.org/pdf/1810.02180.pdf.

Bartlett, P., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. In *NIPS*, pp. 6240–6249, 2017.

Bécigneul, G. On the effect of pooling on the geometry of representations. Technical report, mar 2017. URL http://arxiv.org/abs/1703.06726.

Chafai, D. Singular Values Of Random Matrices. Technical report.

Cullina, D., Bhagoji, A. N., and Mittal, P. PAC-learning in the presence of evasion adversaries. In *NIPS*, 2018.

Fawzi, A., Fawzi, O., and Frossard, P. Analysis of classifiers' robustness to adversarial perturbations. *Mach. Learn.*, 107 (3):481–508, 2018. doi: 10.1007/s10994-017-5663-3.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

Glorot, X., Bordes, A., and Bengio, Y. Deep Sparse Rectifier Neural Networks. In *AISTATS*, 2011.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and Harnessing Adversarial Examples. In *ICLR*, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Identity Mappings in Deep Residual Networks. In *ECCV*, 2016.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. In *NeuIPS*, pp. 125–136, 2019.

ImageNet, T. Tiny imagenet, 2018. URL https://tiny-imagenet.herokuapp.com/.

Jia, K., Li, S., Wen, Y., Liu, T., and Tao, D. Orthogonal Deep Neural Networks. Technical report, 2019. URL http://arxiv.org/abs/1905.05929.

Kannan, H., Kurakin, A., and Goodfellow, I. Adversarial logit pairing. Technical report, 2018. URL http://arxiv.org/abs/1803.06373.

Khim, J. and Loh, P.-L. Adversarial Risk Bounds for Binary Classification via Function Transformation. Technical report, 2018. URL http://arxiv.org/abs/1810.09519.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial Machine Learning at Scale. In *ICLR*, 2017.

Lee, C.-y., Xie, S., and Gallagher, P. W. Deeply-Supervised Nets. In *AISTATS*, 2015.

Li, Y., Min, M. R., Yu, W., Hsieh, C.-J., Lee, T. C. M., and Kruus, E. Optimal Transport Classifier: Defending Against Adversarial Attacks by Regularized Deep Embedding. Technical report, 2018. URL http://arxiv.org/abs/1811.07950.

Lyu, C., Huang, K., and Liang, H. N. A unified gradient regularization family for adversarial examples. In *ICDM*, 2015.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*, 2018.

Miyato, T., Maeda, S. I., Ishii, S., and Koyama, M. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *PAMI*, pp. 1–16, 2018. ISSN 19393539. doi: 10.1109/TPAMI.2018.2858821.

Moosavi-Dezfooli, Mohsen, S., Fawzi, A., Uesato, J., and Frossard, P. Robustness via curvature regularization, and vice versa. In *CVPR*, pp. 9070–9078, 2019.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. In *NeuIPS*, pp. 11615–11626, 2019.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks. In *ICLR*, 2018a.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. The Role Of Over-parametrization In Generalization Of Neural Networks. In *ICLR*, 2018b.

Pfeiffer, F. W. Automatic differentiation in prose. In *ICLR Workshop*, 2017.

Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., and Kohli, P. Adversarial Robustness through Local Linearization. In *NeurIPS*, pp. 13847–13856, 2019.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. M. Adversarially Robust Generalization Requires More Data. In *NIPS*, pp. 5014–5026, 2018.

Sedghi, H., Gupta, V., and Long, P. M. The Singular Values of Convolutional Layers. In *ICLR*, 2018.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014. ISBN 9781107057135.

Sinha, A., Namkoong, H., and Duchi, J. Certifying Some Distributional Robustness with Principled Adversarial Training. In *ICLR*, 2018.

Sokolic, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. D. Robust Large Margin Deep Neural Networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, aug 2017. ISSN 1053-587X. doi: 10.1109/TSP.2017.2708039.

Su, D., Zhang, H., Chen, H., Yi, J., and Aug, C. V. Is Robustness the Cost of Accuracy ? – A Comprehensive Study on the Robustness of. In *ECCV*, 2018.

Szegedy, C., Zaremba, W., and Sutskever, I. Intriguing properties of neural networks. In *ICLR*, 2014.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. Technical report, 2017. URL http://arxiv.org/abs/1705.07204.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Mądry, A. Robustness May Be at Odds with Accuracy. In *ICLR*, 2019.

Tsuzuku, Y., Sato, I., and Sugiyama, M. Lipschitz-Margin Training : Scalable Certification of Perturbation Invariance for Deep Neural Networks. In *NIPS*, pp. 6541–6550, 2018.

Verma, N. Distance Preserving Embeddings for General n-Dimensional Manifolds. *Journal of Machine Learning Research*, 14:2415–2448, 2013.

Wang, J. and Zhang, H. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In *ICCV*, pp. 6629–6638, 2019.

Xie, C., Wu, Y., Maaten, L. v. d., Yuille, A. L., and He, K. Feature denoising for improving adversarial robustness. In *CVPR*, pp. 501–509, 2019.

Xu, H. and Mannor, S. Robustness and generalization. *Machine Learning*, 86(3):391–423, 2012. ISSN 08856125. doi: 10.1007/s10994-011-5268-1.

Yin, D. and Bartlett, P. Rademacher Complexity for Adversarially Robust. In *ICML*, 2018.

Zagoruyko, S. and Komodakis, N. Wide Residual Networks. In *BMVC*, 2016.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *ICLR*, 2016.

Zhang, H., Dauphin, Y. N., and Ma, T. Fixup Initialization: Residual Learning Without Normalization. In *ICLR*, 2018.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically Principled Trade-off between Robustness and Accuracy. In *ICML*, 2019.