Incorporating Label Uncertainty in Intrinsic Robustness Measures

Xiao Zhang and David Evans
Department of Computer Science
University of Virginia
{shawn, evans}@virginia.edu

Abstract

Starting with Gilmer et al. (2018), a line of theoretical works have focused on studying the concentration of measure phenomenon which is fundamentally connected to adversarial robustness. In this work, we argue that the standard concentration is not sufficient to characterize the intrinsic robustness limit for an adversarially robust classification problem since it does not take data labels into account. Built upon on a novel definition of label uncertainty, we empirically demonstrate that error regions induced by various state-of-the-art classification models tend to have much higher label uncertainty than randomly selected subsets. This observation implies that in order to obtain a more accurate intrinsic robustness limit for a particular data distribution, it is important to understand the concentration of measure regarding the input regions with high label uncertainty. In this paper, we adapt the standard concentration problem to produce a more accurate estimate of intrinsic robustness that incorporates label uncertainty and study the error region characteristics of the state-of-the-art machine learning classifiers.

1 Introduction

Since the initial reports of adversarial examples against DNNs (Szegedy et al., 2014; Goodfellow et al., 2015), many defensive mechanisms have been proposed aiming to enhance the robustness of machine learning classifiers, but most have failed against stronger adaptive attacks (Athalye et al., 2018; Tramer et al., 2020). PGD-based adversarial training (Mądry et al., 2018) and its variants (Zhang et al., 2019; Carmon et al., 2019) are among the few heuristic defenses that have not been broken so far, but these methods still fail to produce satisfactorily robust classifiers, even for classification tasks on benchmark datasets like CIFAR-10. Motivated by the empirical hardness of adversarially-robust learning, a line of theoretical works (Gilmer et al., 2018; Fawzi et al., 2018; Mahloujifar et al., 2019a; Shafahi et al., 2019) have argued that adversarial examples are unavoidable. In particular, these works proved that as long as the input distributions are concentrated with respect to the perturbation metric, adversarially robust classifiers do not exist. Recently, Mahloujifar et al. (2019b) and Prescott et al. (2021) generalized these results by developing empirical methods for measuring the concentration of arbitrary input distribution that can be further translated into an intrinsic robustness limit. (See Appendix A for a more complete discussion of related work.)

In this work, we argue that the standard concentration of measure problem, which was studied in all of the aforementioned theoretical works, is not sufficient to capture a realistic intrinsic robustness limit for an adversarially robust classification problem. In particular, the standard concentration function is defined as an inherent property regarding the input metric probability space, which does not take account of the underlying label information. However, such label information is an essential component for any supervised learning problem, including adversarially robust classification.

Contributions. We identify the insufficiency of the standard concentration of measure problem and provide explanations on why it fails to capture a realistic intrinsic robustness limit (Section 2). Then, we introduce the notion of label uncertainty (Definition 3.1), which characterizes the average uncertainty level of the underlying label assignments for an input region. Built upon this definition, we propose to incorporate label uncertainty in the standard concentration measure as an initial step towards a more realistic characterization of intrinsic robustness (Section 3). Experiments on the

CIFAR-10 and CIFAR-10H datasets (Peterson et al., 2019) demonstrate that error regions induced by a wide range of state-of-the-art classification models all have high label uncertainty, which validates the proposed label uncertainty constrained concentration problem (Section 4).

2 Standard Concentration is not Sufficient

In this section, we first explain a fundamental connection between the concentration of measure problem and the intrinsic robustness with respect to imperfect classifiers shown in previous work, and then argue that standard concentration fails to capture a realistic intrinsic robustness limit because it ignores the actual data labels. Appendix B provides formal definitions of the adversarial risk, the intrinsic robustness and the concentration function.

Connecting Intrinsic Robustness with Concentration of Measure. Let $(\mathcal{X}, \mu, \Delta)$ be the considered input metric probability space, $\mathcal{Y}$ be the set of possible labels, and $c : \mathcal{X} \rightarrow \mathcal{Y}$ be the concept function that gives each input a label. Given parameters $0 < \alpha < 1$ and $\epsilon \geq 0$, the standard concentration problem can be cast into an optimization problem as follows:

$$\underset{\mathcal{E}\in\mathsf{pow}(\mathcal{X})}{\text{minimize}}\ \mu(\mathcal{E}_\epsilon) \quad \text{subject to} \quad \mu(\mathcal{E}) \geq \alpha. \tag{2.1}$$

For any classifier f, let $\mathcal{E}_f = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \neq c(\boldsymbol{x})\}$ be its induced error region with respect to $c(\cdot)$. By connecting the risk of f with the measure of $\mathcal{E}_f$ and the adversarial risk of f with the measure of the ϵ-expansion of $\mathcal{E}_f$, Mahloujifar et al. (2019a) proved that the standard concentration problem (2.1) is equivalent to the following optimization problem regarding risk and adversarial risk:

$$\underset{f}{\text{minimize}}\ \mathrm{AdvRisk}_\epsilon(f, c) \quad \text{subject to} \quad \mathrm{Risk}(f, c) \geq \alpha.$$

More specifically, the following lemma characterizes the fundamental connection between the concentration function and the intrinsic robustness with respect to the set of imperfect classifiers:

Lemma 2.1 (Mahloujifar et al. (2019a))**.** For any $\alpha \in (0, 1)$, let $\mathcal{F}_\alpha = \{f : \mathrm{Risk}(f, c) \geq \alpha\}$ be the set of imperfect classifiers, then it holds for any $\epsilon \geq 0$ that

$$\overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}_\alpha, c) = 1 - h(\mu, \alpha, \epsilon).$$

Lemma 2.1 suggests that the concentration function of the input metric probability space $h(\mu, \alpha, \epsilon)$ can be translated into an adversarial robustness upper bound that applies to any classifier with risk at least α. If this upper bound is shown to be small, then one can conclude that it is impossible to learn an adversarially robust classifier, as long as the learned classifier has risk at least α.

Insufficiency of Standard Concentration. Despite of the appealing relationship between concentration of measure and intrinsic robustness, we argue that solving the standard concentration problem may not be sufficient to capture a meaningful intrinsic limit for adversarially robust classification. Recall that the standard concentration of measure problem (2.1) aims to find the optimal subset that has the smallest ϵ-expansion with regard to the input metric probability space $(\mathcal{X}, \mu, \Delta)$, whereas the underlying concept function $c(\cdot)$ that determines the underlying class label of each input is not involved. That said, for the considered metric probability space, no matter how we assign the labels to the inputs, the concentration function $h(\mu, \alpha, \epsilon)$ will remain the same. In sharp contrast, learning an adversarially-robust classifier relies on the joint distribution of both the inputs and the labels.

Moreover, when the standard concentration function is translated into an upper bound on adversarial robustness, it is defined with respect to the set of imperfect classfiers $\mathcal{F}_\alpha$ (see Lemma 2.1). Note that the only restriction imposed by $\mathcal{F}_\alpha$ is that the risk of the classifier (or equivalently, the measure of the corresponding error region) is at least α. Note that, unlike adversarially robust learning, this does not consider whether the classifier is learnable. Thus, it is very likely that the optimal classifier implied by the standard concentration function cannot be produced by any supervised learning method. That said, suppose $\mathcal{F}_{\text{learn}}$ denotes the set of learnable classifiers, then $\overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}_\alpha, c)$ could be much higher than $\overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}_{\text{learn}}, c)$. In fact, it has been observed in Mahloujifar et al. (2019b) that the intrinsic robustness limit implied by concentration of measure is much higher than the adversarial robustness attained by state-of-the-art robust training methods on several image benchmarks.

3 Incorporating Label Uncertainty in Intrinsic Robustness

In this section, we introduce our definition of label uncertainty and how to include it in the intrinsic robustness measures. Let $(\mathcal{X}, \mu)$ be the input probability space and $\mathcal{Y} = \{1, 2, \ldots, k\}$ denote the complete set of labels. A function $\eta : \mathcal{X} \rightarrow [0, 1]^k$ is said to capture the *full label distribution* (Geng, 2016; Gao et al., 2017), if $[\eta(\boldsymbol{x})]_y$ corresponds to the description degree of y to $\boldsymbol{x}$ for any $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, and $\sum_{y=1}^{k} [\eta(\boldsymbol{x})]_y = 1$ holds for any $\boldsymbol{x} \in \mathcal{X}$. For classification tasks that rely on human labeling, one can approximate the label distribution with respect to any input by collecting human labels from multiple human annotators (Peterson et al., 2019).

For any subset $\mathcal{E} \in \mathsf{pow}(\mathcal{X})$, we introduce the notion of *label uncertainty* to capture the average uncertainty level with respect to the label assignments of the inputs within $\mathcal{E}$:

Definition 3.1 (Label Uncertainty)**.** Let $(\mathcal{X}, \mu)$ be the input probability space and $\mathcal{Y} = \{1, 2, \ldots, k\}$ be the complete set of class labels. Suppose $c : \mathcal{X} \rightarrow \mathcal{Y}$ is a concept function that assigns each input $\boldsymbol{x}$ a label $y \in \mathcal{Y}$. Assume $\eta : \mathcal{X} \rightarrow [0, 1]^k$ is the underlying label distribution function, where $[\eta(\boldsymbol{x})]_y$ represents the description degree of y to $\boldsymbol{x}$. For any subset $\mathcal{E} \in \mathsf{pow}(\mathcal{X})$ with measure $\mu(\mathcal{E}) > 0$, the *label uncertainty* of $\mathcal{E}$ with respect to $(\mathcal{X}, \mu)$, $c(\cdot)$ and $\eta(\cdot)$ is defined as:

$$\mathrm{LU}(\mathcal{E}; \mu, c, \eta) = \frac{1}{\mu(\mathcal{E})} \int_{\mathcal{E}} \Big\{ 1 - \big[\eta(\boldsymbol{x})\big]_{c(\boldsymbol{x})} + \max_{y' \neq c(\boldsymbol{x})} \big[\eta(\boldsymbol{x})\big]_{y'} \Big\} \, d\mu.$$

The range of label uncertainty is $[0, 2]$. We define $\mathrm{LU}(\mathcal{E}; \mu, c, \eta)$ as the average label uncertainty for all the examples that fall into $\mathcal{E}$, where $1 - [\eta(\boldsymbol{x})]_{c(\boldsymbol{x})} + \max_{y' \neq c(\boldsymbol{x})} [\eta(\boldsymbol{x})]_{y'}$ represents the label uncertainty of a single example $\{\boldsymbol{x}, c(\boldsymbol{x})\}$. For a single input, label uncertainty of 0 suggests the assigned label fully captures the underlying label distribution, label uncertainty of 1 means there are other classes as likely to be the ground-truth label as the assigned label, and label uncertainty of 2 means there is a label other than the assigned one that should be the ground-truth.

State-of-the-art classification models are expected to misclassfy more inputs with large label uncertainty, as there is more discrepancy between their assigned labels and the underlying label distribution (see Section 4 for empirical evidence on CIFAR-10). Thus, to obtain a more realistic intrinsic robustness limit, we propose the following label uncertainty constrained concentration problem:

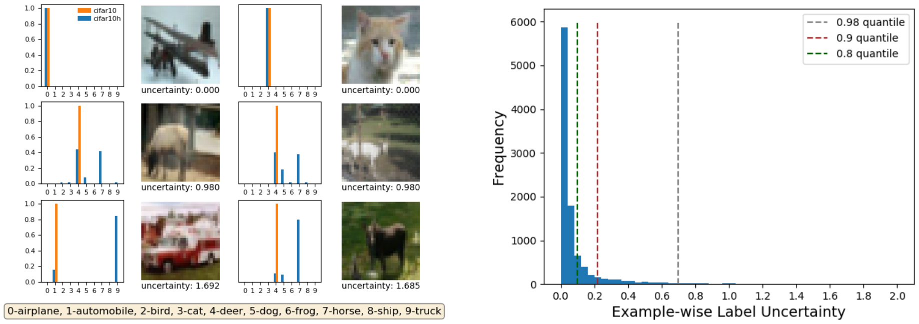
$$\underset{\mathcal{E} \in \mathsf{pow}(\mathcal{X})}{\text{minimize}} \ \mu(\mathcal{E}_\epsilon) \quad \text{subject to} \quad \mu(\mathcal{E}) \geq \alpha \ \text{ and } \ \mathrm{LU}(\mathcal{E}; \mu, c, \eta) \geq \gamma, \tag{3.1}$$

where $\gamma \in [0, 2]$ is a constant that will be determined based on the given classification problem. Note that when γ is set as zero, (3.1) degenerates to the standard concentration of measure problem. Similar to the connection between the standard concentration function and $\overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}_\alpha)$, the optimal value of (3.1) is equivalent to $1 - \overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}_{\alpha,\gamma})$, where $\mathcal{F}_{\alpha,\gamma}$ denotes the set of classifiers with risk at least α and error region label uncertainty at least γ. As long as a classifier f belongs to $\mathcal{F}_{\alpha,\gamma}$, it is guaranteed that $\mathrm{AdvRob}_\epsilon(f) \leq \overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}_{\alpha,\gamma}) \leq \overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}_\alpha)$. Although both $\overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}_\alpha)$ and $\overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}_{\alpha,\gamma})$ can serve as valid robustness upper bounds, the latter one would be able to capture a more meaningful intrinsic robustness limit, since it takes account of the fact that state-of-the-art classification models prone to misclassify inputs with higher label uncertainty.

4 Experimental Results

To illustrate our label uncertainty definition and test the aforementioned hypothesis, we conduct experiments on the CIFAR-10H dataset (Peterson et al., 2019), which contains soft labels reflecting human perceptual uncertainty for the 10,000 CIFAR-10 test images. These soft labels can be regarded as an approximation of the label distribution function $\eta(\cdot)$ at each given input, whereas the original CIFAR-10 test dataset provides the class labels given by the concept function $c(\cdot)$.

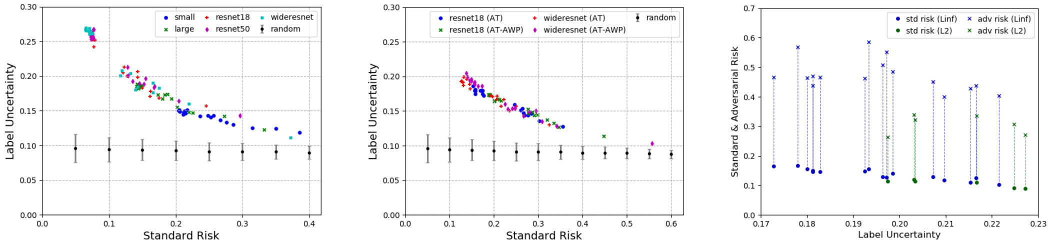
Figure 1(a) shows the label uncertainty score for several images with both the soft labels from CIFAR-10H and the original class labels from CIFAR-10. Images with low uncertainty scores are typically easier for humans to recognize their class category (e.g., the first row of Figure 1(a)), whereas images with high uncertainty scores look ambiguous or even misleading (e.g., the second and third rows). Figure 1(b) shows the histogram of the label uncertainty distribution for all the CIFAR-10 test examples. More than 80% examples have label uncertainty score less than 0.1, suggesting their



(a) Illustration of CIFAR-10 and CIFAR-10H

(b) Label Uncertainty Distribution

Figure 1: (a) Visualization of the CIFAR-10 test images with the soft labels from CIFAR-10H, the original assigned labels from CIFAR-10 and the label uncertainty scores computed based on Definition 3.1. (b) Histogram of the label uncertainty distribution for the CIFAR-10 test dataset.



(a) Standard Training

(b) Adversarial Training

(c) RobustBench

Figure 2: Visualizations of error region label uncertainty versus standard risk and adversarial risk with respect to classifiers produced by different machine learning methods: (a) Standard-trained classifiers with different network architecture; (b) Adversarially-trained classifiers using different learning algorithms; (c) State-of-the-art adversarially robust classification models from RobustBench.

original class labels do capture the underlying label distribution well. However, around 2% of the examples have label uncertainty scores exceeding 0.7.

We hypothesize that ambiguous or misleading images should also be more likely to be misclassified as errors by state-of-the-art machine learning classifiers, or in other words, their induced error regions should have larger label uncertainty. In order to test this hypothesis, we conduct experiments on CIFAR-10 and CIFAR-10H datasets. More specifically, we train different classification models, including intermediate models extracted at different epochs, using the CIFAR-10 training dataset, then empirically compute the standard risk, the adversarial risk and the label uncertainty of the corresponding error region. The results are shown in Figure 2 (see Appendix C for model specifics and training hyperparameters).

Figures 2(a) and 2(b) demonstrate the relationship between label uncertainty and standard risk for various classifiers produced by standard training method and adversarial training methods under ℓ_∞ perturbations with $\epsilon = 8/255$. In addition, we plot the label uncertainty with error bars of randomly-selected images from the CIFAR-10 test dataset as a reference. As the model classification accuracy increases, the label uncertainty of its induced error region increases, suggesting the misclassified examples tend to have higher label uncertainty. This observation holds consistently for both standard and adversarially trained models with any tested network architecture. Figure 2(c) summarize sthe error region label uncertainty with respect to the state-of-the-art adversarially robust models documented in RobustBench (Croce et al., 2020). Regardless of the perturbation type or the learning method, the average label uncertainty of their misclassified examples all falls into a range of $(0.17, 0.23)$, whereas the mean label uncertainty of all the testing CIFAR-10 data is less than 0.1. This validates our hypothesis that error regions of state-of-the-art classifiers tend to have larger label uncertainty, and supports the need for accounting for labels in intrinsic robustness.

References

Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? In *NeurIPS*, 2019.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 2018.

Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In *NeurIPS*, 2019.

Christer Borell. The Brunn-Minkowski inequality in Gauss space. *Inventiones mathematicae*, 30(2): 207–216, 1975.

Sebastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, 2019.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *NeurIPS*, 2019.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Dimitrios Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. Adversarial risk and robustness: General definitions and implications for the uniform distribution. In *NeurIPS*, 2018.

Elvis Dohmatob. Generalized no free lunch theorem for adversarial robustness. In *International Conference on Machine Learning*, 2019.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. In *NeurIPS*, 2018.

Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017.

Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28 (7):1734–1748, 2016.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv:1801.02774*, 2018.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *IEEE International Conference on Computer Vision*, 2019.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, 2019.

Saeed Mahloujifar, Dimitrios Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. In *AAAI Conference on Artificial Intelligence*, 2019a.

Saeed Mahloujifar, Xiao Zhang, Mohammad Mahmoody, and David Evans. Empirically measuring concentration: Fundamental limits on intrinsic robustness. In *NeurIPS*, 2019b.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, 2016.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *IEEE International Conference on Computer Vision*, 2019.

Jack Prescott, Xiao Zhang, and David Evans. Improved estimation of concentration under ℓ_p-norm distance metrics using half spaces. In *International Conference on Learning Representations*, 2021.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018.

Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? In *International Conference on Learning Representations*, 2019.

Vladimir N Sudakov and Boris S Tsirelson. Extremal properties of half-spaces for spherically invariant measures. *Zapiski Nauchnykh Seminarov Leningrad Otdel Mathematical Institute Steklov* (LOMI), 41:14–24, 1974.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, 81(1):73–205, 1995.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *arXiv:2002.08347*, 2020.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, 2018.

Eric Wong, Frank R Schmidt, Jan Hendrik Metzen, and Zico Kolter. Scaling provable adversarial defenses. In *NeurIPS*, 2018.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. In *International Conference on Learning Representations*, 2020a.

Xiao Zhang, Jinghui Chen, Quanquan Gu, and David Evans. Understanding the intrinsic robustness of image distributions using conditional generative models. In *International Conference on Artificial Intelligence and Statistics* (AISTATS), 2020b.

A Related Work

A.1 Training Adversarially Robust Classifier

Witnessing the vulnerability of modern machine learning models to adversarial examples, extensive studies have been carried out aiming to build classification models that can be robust against adversarial perturbations. Heuristic defense mechanisms (Goodfellow et al., 2015; Papernot et al., 2016; Guo et al., 2018; Xie et al., 2018; Mądry et al., 2018) had been most popular until many of them were broken by stronger adaptive adversaries (Athalye et al., 2018; Tramer et al., 2020). The only exception, which has not been defeated yet so far, is PGD-based adversarial training (Mądry et al., 2018). Recent successful methods are variants of PGD-based adversarial training, which either adopt different loss function (Zhang et al., 2019; Wu et al., 2020) or make use of additional training data (Carmon et al., 2019; Alayrac et al., 2019). Nevertheless, the best current adversarially trained classifiers can only achieve around 60% robust accuracy on CIFAR-10 against ℓ_∞ perturbations with strength $\epsilon = 8/255$, even with additional training data (see the leaderboard in Croce et al. (2020)).

In addition, to end the arms race between heuristic defenses and newly designed adaptive attacks that break them, certified defenses have been developed based on different approaches, including linear programming (Wong & Kolter, 2018; Wong et al., 2018), semidefinite programming (Raghunathan et al., 2018), interval bound propagation (Gowal et al., 2019; Zhang et al., 2020a) and randomized smoothing (Cohen et al., 2019; Li et al., 2019). Although certified defenses are able to train classifiers with robustness guarantees for input instances, they usually come with sacrificed empirical robustness, especially for larger adversarial perturbations.

A.2 Theoretical Understanding on Intrinsic Robustness

Given the unsatisfactory status quo of building adversarially robust classification models, a line of research (Gilmer et al., 2018; Fawzi et al., 2018; Mahloujifar et al., 2019a; Shafahi et al., 2019; Dohmatob, 2019; Bhagoji et al., 2019) attempted to explain the adversarial vulnerability from a theoretical perspective. To be more specific, they proved that if the input distribution is concentrated with respect to the perturbation metric, then there does not exist adversarially robust classifiers. At the core of these results is the fundamental connection between the concentration of measure phenomenon and an intrinsic robustness limit, capturing the maximum adversarial robustness with respect to some specific set of classifiers. For instance, Gilmer et al. (2018) showed that for inputs sampled from uniform n-spheres, a model-independent robustness upper bound under Euclidean distance metric can be derived using Gaussian Isoperimetric Inequality (Sudakov & Tsirelson, 1974; Borell, 1975), whereas Mahloujifar et al. (2019a) generalized their result to any concentrated metric probability space of inputs. Nevertheless, it still remains unclear how to apply these theoretical results to typical image classification tasks, since whether or not natural image distributions are concentrated is unknown.

In order to address such question, Mahloujifar et al. (2019b) proposed a general way to measure the concentration for any input distribution using data samples, then employed it to estimate an intrinsic robustness limit for typical image benchmarks. By showing the existence of a large gap between the limit implied by concentration and the empirical robustness achieved by state-of-the-art adversarial training methods, Mahloujifar et al. (2019b) further concluded that concentration of measure can only explain a small portion of adversarial vulnerability of existing image classifiers. More recently, Prescott et al. (2021) further strengthened their conclusion by using the set of half-spaces to estimate the concentration function, which achieves enhanced estimation accuracy.

B Preliminaries

This section introduces the most related topics, including the definition of adversarial risk, the definition of intrinsic robustness, and the standard concentration of measure problem.

Notation. We use lowercase boldface letters such as $\boldsymbol{x}$ to denote vectors. For any set $\mathcal{A}$, let $\mathsf{pow}(\mathcal{A})$ and $\mathbb{1}_{\mathcal{A}}(\cdot)$ be all measurable subsets and the indicator function of $\mathcal{A}$. Consider metric probability space $(\mathcal{X}, \mu, \Delta)$, where $\Delta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is a distance metric on $\mathcal{X}$. Denote by $\mathcal{B}_\epsilon^{(\Delta)}(\boldsymbol{x})$ the ball around $\boldsymbol{x}$

with radius ϵ measured by Δ. Define the ϵ-expansion of $\mathcal{A}$ as $\mathcal{A}_\epsilon^{(\Delta)} = \{\boldsymbol{x} \in \mathcal{X} : \exists \boldsymbol{x}' \in \mathcal{B}_\epsilon^{(\Delta)}(\boldsymbol{x}) \cap \mathcal{A}\}$. When Δ is free of context, we simply write $\mathcal{B}_\epsilon(\boldsymbol{x}) = \mathcal{B}_\epsilon^{(\Delta)}(\boldsymbol{x})$ and $\mathcal{A}_\epsilon = \mathcal{A}_\epsilon^{(\Delta)}$.

Adversarial Risk. Adversarial risk captures the vulnerability of a classifier against adversarial perturbations. In particular, we work with the following adversarial risk definition[1], which has been studied in several previous works, such as Gilmer et al. (2018); Bubeck et al. (2019); Mahloujifar et al. (2019a;b); Zhang et al. (2020b); Prescott et al. (2021).

Definition B.1 (Adversarial Risk)**.** Let $(\mathcal{X}, \mu, \Delta)$ be a metric probability space of instances and $\mathcal{Y}$ be the set of possible class labels. Assume $c : \mathcal{X} \to \mathcal{Y}$ is a concept function that gives each instance a label. For any classifier $f : \mathcal{X} \to \mathcal{Y}$ and $\epsilon \geq 0$, the *adversarial risk* of f is defined as:

$$\mathrm{AdvRisk}_\epsilon(f, c) = \Pr_{\boldsymbol{x} \sim \mu}\big[\exists\, \boldsymbol{x}' \in \mathcal{B}_\epsilon(\boldsymbol{x}) \text{ s.t. } f(\boldsymbol{x}') \neq c(\boldsymbol{x}')\big].$$

The *adversarial robustness*[2] of f is defined as: $\mathrm{AdvRob}_\epsilon(f, c) = 1 - \mathrm{AdvRisk}_\epsilon(f, c)$.

When $\epsilon = 0$, adversarial risk equals to the standard risk. Namely, $\mathrm{AdvRisk}_0(f, c) = \mathrm{Risk}(f, c) := \Pr_{\boldsymbol{x} \sim \mu}[f(\boldsymbol{x}) \neq c(\boldsymbol{x})]$ holds for any classifier f. It is worth noting that other related definitions of adversarial risk, such as the one used in most empirical works for robustness evaluation, are equivalent to ours, as long as small perturbations do not change the labels assigned by $c(\cdot)$.

Intrinsic Robustness. The definition of intrinsic robustness was first introduced by Mahloujifar et al. (2019b) to capture the maximum adversarial robustness with respect to some set of classifiers:

Definition B.2 (Intrinsic Robustness)**.** Consider the input metric probability space $(\mathcal{X}, \mu, \Delta)$ and the set of labels $\mathcal{Y}$. Let $c : \mathcal{X} \to \mathcal{Y}$ be a concept function that gives a label to each input. For any set of classifiers $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathcal{Y}\}$ and $\epsilon \geq 0$, the *intrinsic robustness* with respect to $\mathcal{F}$ is defined as:

$$\overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}, c) = 1 - \inf_{f \in \mathcal{F}} \big\{\mathrm{AdvRisk}_\epsilon(f, c)\big\} = \sup_{f \in \mathcal{F}} \{\mathrm{AdvRob}_\epsilon(f, c)\}.$$

According to the definition of intrinsic robustness, one immediately know that there does not exist any classifier in $\mathcal{F}$ with adversarial robustness higher than $\overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}, c)$ for the considered adversarially robust classification task. Existing works, including Gilmer et al. (2018); Mahloujifar et al. (2019a;b); Zhang et al. (2020b), selected $\mathcal{F}$ in Definition B.2 as the set of imperfect classifiers $\mathcal{F}_\alpha = \{f : \mathrm{Risk}(f, c) \geq \alpha\}$, where $\alpha \in (0, 1)$ is some constant.

Concentration of Measure. Concentration of measure captures a 'closeness' property for a metric probability space of instances. More formally, it is defined by the concentration function as follows:

Definition B.3 (Concentration Function)**.** Let $(\mathcal{X}, \mu, \Delta)$ be a metric probability space. For any $0 < \alpha < 1$ and $\epsilon \geq 0$, the *concentration function* of $(\mathcal{X}, \mu, \Delta)$ is defined as:

$$h(\mu, \alpha, \epsilon) = \inf_{\mathcal{E} \in \mathsf{pow}(\mathcal{X})} \{\mu(\mathcal{E}_\epsilon) : \mu(\mathcal{E}) \geq \alpha\}.$$

The standard notion of concentration function considers a special case of Definition B.3 with $\alpha = 1/2$ (e.g., see Talagrand (1995)). For some special metric probability spaces, one can prove the closed-form solution of the concentration function. For instance, Gaussian Isoperimetric Inequality (Borell, 1975; Sudakov & Tsirelson, 1974) characterizes the concentration function for spherical Gaussian distribution and ℓ_2-norm distance metric.

C Detailed Experimental Settings

This section specifies the details of the experiments presented in Section 4.

[1] There exist other definitions of adversarial risk in literature, such as the one used in Mądry et al. (2018). However, these definitions are equivalent ours, as long as small perturbation preserves the ground truth.

[2] The average distance to the error region has also been used in literature, such as Diochnos et al. (2018), as an indicator of adversarial robustness. Both of these definitions characterize how resilient the classifier is against adversarial perturbations.

For standard trained classifiers, we implemented five neural network architecture, including a 4-layer neural net with two convolutional layers and two fully-connected layers (*small*), a 7-layer neural net with four convolutional layers and three fully-connected layers (*large*), a ResNet-18 architecture (*resnet18*), ResNet-50 architecture (*resnet50*) and a WideResNet-34-10 architecture (*wideresnet*). We trained the *small* and *large* model using a Adam optimizer with initial learning rate 0.005, whereas we trained the *resnet18*, *resnet50* and *wideresnet* model using a SGD optimizer with initial learning rate 0.01. All models are trained using a piece-wise learning rate schedule with a decaying factor of 10 at epoch 50 and epoch 75, respectively. For Figure 2(a), we plotted the label uncertainty and standard risk for the intermediate models obtained at epochs $5, 10, \ldots, 100$ for each architecture. In addition, we also randomly selected different subsets of inputs with empirical measure of $0.05, 0.10, \ldots 0.95$ and plotted their corresponding label uncertainty with error bars.

For adversarially trained classifiers, we implemented the vanilla adversarial training method (Mądry et al., 2018) and the adversarial training method with adversarial weight perturbation (Wu et al., 2020), which are denoted as *AT* and *AT-AWP* in Figure 2(b) respectively. Both ResNet-18 (*resnet18*) and WideResNet-34-10 (*wideresnet*) architecture are implemented for each training method. A 10-step PGD attack (PGD-10) with step size $2/255$ and maximum perturbation size $8/255$ is used for each model during training. In addition, each model is trained for 200 epochs using a SGD optimizer with initial learning rate 0.1 and piece-wise learning rate schedule with a decaying factor of 10 at epoch 100 and epoch 150. We record the intermediate models at epoch $10, 20, \ldots, 200$ respectively.