From Characterizing Intrinsic Robustness to Adversarially Robust Machine Learning

---

A

Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Xiao Zhang

August 2022

APPROVAL SHEET

This

Dissertation

is submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Author: Xiao Zhang

This Dissertation has been read and approved by the examing committee:

Advisor: David Evans

Advisor:

Committee Member: Tom Fletcher

Committee Member: Mohammad Mahmoody

Committee Member: David Wu

Committee Member: Somesh Jha

Committee Member: Tianxi Li

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

August 2022

Abstract

The prevalence of adversarial examples raises questions about the reliability of machine learning systems, especially for their deployment in critical applications. Numerous defense mechanisms have been proposed that aim to improve a machine learning system's robustness in the presence of adversarial examples. However, none of these methods are able to produce satisfactorily robust models, even for simple classification tasks on benchmarks. In addition to empirical attempts to build robust models, recent studies have identified intrinsic limitations for robust learning against adversarial examples. My research aims to gain a deeper understanding of why machine learning models fail in the presence of adversaries and design ways to build better robust systems. In this dissertation, I develop a concentration estimation framework to characterize the intrinsic limits of robustness for typical classification tasks of interest. The proposed framework leads to the discovery that compared with the concentration of measure which was previously argued to be an important factor, the existence of uncertain inputs may explain more fundamentally the vulnerability of state-of-the-art defenses. Moreover, to further advance our understanding of adversarial examples, I introduce a notion of representation robustness based on mutual information, which is shown to be related to an intrinsic limit of model robustness for downstream classification tasks. Finally in this dissertation, I advocate for a need to rethink the current design goal of robustness and shed light on ways to build better robust machine learning systems, potentially escaping the intrinsic limits of robustness.

To my family.

Acknowledgments

Throughout my Ph.D. journey, I am very grateful to receive help from many people. I would like to show my sincere appreciation to all of them who have made me a better person.

First and foremost, my deepest gratitude and appreciation go to my perfect advisor, Dr. David Evans. He has always been a role model to me as an inspiring researcher and a supportive mentor. He taught me how to do great research and always believed in me during my Ph.D. studies, even when I was not sure how to proceed with my research. Without his constant encouragement and guidance, I would have never made up my mind to pursue a faculty career for the rest of my life. His dedication and rigorous attitude in research deeply affected me. He made me wish to become a great mentor like him that can bring knowledge and inspiration to many others.

Second, I would like to express my gratitude to my Ph.D. committee members, Dr. Tom Fletcher, Dr. Mohammad Mahmoody, Dr. David Wu, Dr. Somesh Jha, and Dr. Tianxi Li, for their valuable time spent on my dissertation research. Their insightful comments and constructive suggestions inspired me greatly and helped shape my dissertation research. In addition, I would like to thank Dr. Quanquan Gu, Dr. Simon Du, Dr. Haifeng Xu, Dr. Yuan Tian, Dr. Tianhao Wang, Dr. Yangfeng Ji, and Dr. Tingting Zhang for their help and advice with my research and career development at different stages of my Ph.D. studies.

Moreover, I would like to say thank you to my great collaborators. In particular, I would like to thank Yaodong Yu, an excellent collaborator and a close friend. We exchanged opinions on research and shared experiences. He always encouraged me during my Ph.D. studies. In addition, I want to specifically acknowledge the contributions of my following collaborators: Dr. Mohammad Mahmoody, Dr. Saeed Mahloujifar, Dr. Jinghui Chen, Sicheng Zhu, and

Jack Prescott. This dissertation benefits a lot from the research collaborations with them.

Furthermore, I would like to express my special thank to Zhiqiu Jiang, my dear wife, best friend, and life partner. Sometimes, things didn't work out the way you'd thought they would, but the accompanying of Zhiqiu and her constant encouragement made me believe in myself and survive those moments. Her curiosity and passion for knowledge also greatly influenced me during my Ph.D. journey and beyond. My life had never been that colorful until Zhiqiu became my significant other.

Last but not least, I want to thank my parents for their unconditional love and endless support for me in pursuing my Ph.D. dream.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning has made remarkable breakthroughs in various fields, including computer vision [55] and natural language processing [27], especially when classification accuracy is evaluated. However, state-of-the-art machine learning models have been shown to be extremely vulnerable to classifying inputs, known as adversarial examples [116, 51], that are crafted with targeted but visually-imperceptible perturbations. This phenomenon has raised serious trustworthy concerns for deploying machine learning models in critical applications, such as malware detection [104], face recognition [110] and autonomous vehicles [42].

Since the initial reports of adversarial examples, many defensive mechanisms [93, 11, 48, 21] have been proposed aiming to enhance the robustness of machine learning models. Most have failed, however, against stronger adaptive attacks [5, 118]. PGD-based adversarial training [79] and its variants [129, 17] are the current state-of-the-art, but these methods still fail to produce satisfactorily robust classifiers, even for benchmark classification tasks on image datasets like CIFAR-10.

In addition to empirical attempts to build adversarially robust models, recent studies have identified intrinsic difficulties for learning in the presence of adversarial examples. In particular, a line of research [49, 44, 80, 108] proved that adversarial robustness is unattainable if the underlying distribution is concentrated with respect to the perturbation metric. Although such findings seem discouraging to the goal of developing robust classifiers, there exists a large gap between the problem settings where the inevitability results of adversar-

ial examples are drawn and the typical classification tasks considered by most empirical works. It remains elusive whether these theoretical results apply to actual classification tasks of interest.

Witnessing the empirical bottleneck for improving model robustness and the negative results on adversarial examples, this dissertation aims to develop provable methods for understanding the fundamental causes of the adversarial vulnerability. The proposed methods shrink the gap between analyses of robustness for theoretical distributions and understanding the intrinsic robustness limits for actual datasets of interest. In addition, they provide quantitative estimates that characterize the contribution of each fundamental factor in explaining adversarial vulnerability, which further suggests promising directions for escaping the intrinsic limits of robustness. Moreover, I am going to identify scenarios where the current design goal of a system's robustness is not appropriate, and shed light on potential ways to build better robust machine learning systems.

1.1 Contributions

The main contributions of this thesis are summarized as follows.

1.1.1 Deeper Understanding of Adversarial Robustness

Characterizing Intrinsic Limits on Classifier Robustness. To understand to what extent the inevitability results of adversarial examples apply to typical robust classification tasks of interest, we developed different empirical methods to understand and estimate the intrinsic limits of adversarial robustness translated by the concentration of measure in a series of publications [131, 81, 96] (Chapter 3). In particular, the first method was introduced in

the paper [131], which proves upper bounds on *intrinsic robustness* (see Definition 3.2) based on the assumption that the underlying data is captured by a conditional generative model, then provides empirical estimates of such bounds on generated image distributions (Section 3.3). The second method was initially proposed as a general approach in the paper [81] then improved by the method introduced in the work [96], where empirical estimators are developed to measure the concentration of an arbitrary distribution using data samples, then employed it to estimate a lower bound on intrinsic robustness for image benchmarks (Sections 3.4 and 3.5). Observing a large gap between the estimated bounds of intrinsic robustness and the robustness performance achieved by the best current adversarially trained models, we conclude that different from the conclusion drawn from previous works [49, 44, 80, 108], concentration of measure should not be considered as the main reason behind the adversarial vulnerability of existing classifiers for typical classification tasks on image benchmarks.

Revealing the Importance of Labels in Intrinsic Robustness. The existence of a large gap between intrinsic robustness estimates and robust accuracies attained by state-of-the-art classifiers motivates us to further study the reasons behind this phenomenon. In the paper [133], we show that intrinsic robustness with respect to imperfect classifiers, studied in all of the aforementioned works, is not sufficient to capture a realistic intrinsic robustness limit, because it ignores data labels which are essential to any classification task. Therefore, we argue that it would be more meaningful to incorporate the underlying label information into the definition of intrinsic robustness.

We introduced a novel definition of *label uncertainty* (see Definition 4.3), and empirically observed that error regions induced by state-of-the-art models all tend to have much higher label uncertainty. This observation motivated us to incorporate label uncertainty in the standard concentration measure as an initial step towards a more realistic characterization

of intrinsic robustness. We further adapted a standard concentration estimation algorithm that accounts for label uncertainty. Our results show that the proposed method is able to produce a lower intrinsic robustness limit for image benchmarks than was possible using prior methods that do not consider data labels, suggesting that in addition to concentration of measure, another fundamental cause of the adversarial vulnerability is the existence of examples with high label uncertainty (Chapter 4).

Improved Understanding of Robustness at the Level of Representations. To better understand adversarial robustness, we considered the underlying problem of learning robust representations [135]. In particular, we introduced a notion of *representation vulnerability* based on mutual information (see Definition 5.4), then proposed an unsupervised learning method for obtaining intrinsically robust representations by maximizing the worst-case mutual information between the input and output distributions. Our results demonstrate a strong correlation between model and representation robustness, suggesting the effectiveness of our method as an approach for understanding and measuring achievable adversarial robustness at the level of representations (Chapter 5).

1.1.2 Towards building better robust classifiers

Although understanding intrinsic limits of classifier robustness is scientifically important, our ultimate goal is to design ways to build better machine learning systems. Built upon the current design goal of *overall* adversarial robustness, however, it seems that there is no hope to escape the intrinsic robustness limits. This motivates me to rethink whether overall robustness is the right design goal for building robust machine learning systems. I argue that there exist scenarios where overall robustness is not the appropriate evaluation criterion

for a system's robustness (Chapter 6), which I further explain below.

Cost-Sensitive Robustness. In the paper [132], we argue that from a security perspective, only certain kinds of adversarial misclassifications pose meaningful threats that provide value for potential adversaries, whereas overall robustness places equal emphasis on every possible adversarial transformation. As a simple example, misclassifying a malicious program as benign results in more severe consequences than the reverse. Motivated by this observation, we proposed a general method for adapting certified defenses against norm-bounded perturbations to take into account the potential harm of different adversarial class transformations. In particular, we capture the impact of different adversarial class transformations using a cost matrix. Instead of reducing the overall robust error, we advocate for a more meaningful goal of maximizing the *cost-sensitive robustness* against adversarial examples. Our results showed that the proposed training method is able to produce models with significantly improved cost-sensitive robustness, while maintaining similar clean accuracy for a variety of cost scenarios on typical image benchmarks (Chapter 6.2).

Uncertainty-Aware Robustness. In addition to not considering the underlying goal of adversaries, I argue that overall robustness would also be an unrealistic goal to achieve for uncertain inputs, since the ground-truth labels are inherently probabilistic at those inputs, thus should not be regarded as single-labeled . Inspired by existing literature on conformal classification [102, 3], I propose the notion of *uncertainty-aware robustness* (see Definition 6.1) as a more meaningful metric for evaluating classifier's robustness, which tolerates adversarial class transformations that are aligned with the underlying label information. In addition, by adapting a PGD-attack based algorithm, I design methods to empirically evaluate the uncertainty-aware robustness of a given classifier and demonstrate the superiority of uncertainty-aware robustness for benchmark image classification tasks, especially when a classifier's robustness performance is being assessed on the set of intrinsically uncertain

inputs (Chapter 6.3).

1.2 Dissertation Structure

Chapter 2 reviews the most related literature to this dissertation. In Chapter 3, we develop different methods for measuring concentration, then employ them to estimate intrinsic robustness limits for benchmark classification tasks. In Chapter 4, we identity the insufficiency of standard concentration function in capturing a meaningful intrinsic robustness limit, then study the concentration problem with the consideration of label uncertainty. To better understand adversarial robustness, we study the underlying problem of learning robust representations in Chapter 5. In Chapter 6, we identify scenarios where overall robustness is not the appropriate criterion for measuring a system's robustness performance, and discuss potential ideas to build better robust ML systems. Finally, I conclude my dissertation and discuss open questions in Chapter 7.

Notations. For any $n \in \mathbb{Z}^+$, denote by $[n]$ the set $\{1, 2, \ldots, n\}$. Lowercase boldface letters such as $\boldsymbol{x}$ denote vectors and uppercase boldface letters such as $\mathbf{A}$ represent matrices. For any vector $\boldsymbol{x}$ and $p \in [1, \infty)$, let x_j, $\|\boldsymbol{x}\|_p$ and $\|\boldsymbol{x}\|_\infty$ be the j-th element, the ℓ_p-norm and the ℓ_∞-norm of $\boldsymbol{x}$. For any matrix $\mathbf{A}$, $\mathbf{B}$ is said to be a square root of $\mathbf{A}$ if $\mathbf{A} = \mathbf{BB}$, and the induced matrix p-norm of $\mathbf{A}$ is defined as $\|\mathbf{A}\|_p = \sup_{\boldsymbol{x} \neq \mathbf{0}} \{\|\mathbf{A}\boldsymbol{x}\|_p / \|\boldsymbol{x}\|_p\}$.

For any set $\mathcal{A}$, $|\mathcal{A}|$ denotes its cardinality, $\text{Pow}(\mathcal{A})$ is all its measurable subsets, and $\mathbb{1}_{\mathcal{A}}(\cdot)$ is 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 $\mathcal{X}$. Define the empirical measure of μ with respect to a data set $\mathcal{S}$ sampled from μ as $\hat{\mu}_{\mathcal{S}}(\mathcal{A}) = \sum_{\boldsymbol{x} \in \mathcal{S}} \mathbb{1}_{\mathcal{A}}(\boldsymbol{x}) / |\mathcal{S}|$. Denote by $\mathcal{B}_\epsilon^{(\Delta)}(\boldsymbol{x})$ the ball around $\boldsymbol{x}$ with radius ϵ measured by Δ. The ϵ-expansion of $\mathcal{A}$ is defined 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 $\mathcal{A}_\epsilon = \mathcal{A}_\epsilon^{(\Delta)}$. The collection of the ϵ-expansions for members of any $\mathcal{G} \subseteq \mathsf{Pow}(\mathcal{X})$ is defined and denoted as $\mathcal{G}_\epsilon = \{\mathcal{A}_\epsilon : \mathcal{A} \in \mathcal{G}\}$.

We use $\mathbf{I}_n$ to denote the $n \times n$ identity matrix. Denote by $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ the Gaussian distribution with mean $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Sigma}$. Let γ_n be the probability measure of $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. For the one dimensional case, we use $\Phi(x)$ to denote the cumulative distribution function (CDF) of $\mathcal{N}(0, 1)$, and use $\Phi^{-1}(x)$ to denote its inverse function. For any function $g : \mathcal{Z} \to \mathcal{X}$ and probability measure ν defined over $\mathcal{Z}$, $g_*(\nu)$ denotes the push-forward measure of ν.

Chapter 2

Related Work

I review the most related literature in this chapter, including both the empirical works that propose defenses against adversarial examples and theoretical works on explaining the hardness of adversarially robust learning.

2.1 Defenses against Adversarial Examples

Since the first discovery of adversarial examples by [116], numerous heuristic defense mechanisms have been proposed aiming to improve model robustness, such as defensive distillation [93], input transformations [53], thermometer encoding [15], randomization schemes [125, 29] and adversarial training [51, 71, 79]. However, many of these proposed methods have been shown to be ineffective against stronger adaptive adversaries [16, 5, 118]. To end this arms race between heuristic defenses that claim successful against existing attacks and newly devised stronger attacks that penetrate these models, various certifiable methods have been proposed, such as linear relaxations [121, 35], semi-definite programming [97], distributionally robust optimization [112] and interval bound propagation [82, 52]. Although these methods are able to certify the classifier's prediction to be constant within some ℓ_p-norm bounded set around given inputs and to train models to optimize for certifiable robustness, they suffer from scalability issues when the size of the datasets or the implemented networks are large, and the best achieved certifiable robustness guaran-

tee is far from satisfying. Despite continuous efforts seeking to develop scalable methods for improving robustness against adversarial examples [120, 129, 130, 19], state-of-the-art training methods cannot produce adversarially robust models, even for simple classification tasks on CIFAR-10 [68]. In this thesis, we are going to explore better ways to build better robust machine learning systems, by rethinking the current design goal of robustness.

2.2 Theoretical works on Adversarially Robust Learning

Given the unsatisfying status of the existing adversarial defenses, recent works attempted to provide fundamental explanations for the difficulties of robust learning against adversarial examples. In particular, one line of research [49, 44, 80, 108, 31, 10] studied the intrinsic robustness limits of robust learning with respect to the metric probability space of the inputs. They showed that no classifier is able to achieve adversarial robustness, if the input distribution and the perturbation set satisfy certain assumptions. For example, the pioneering work of [49] proved model-independent bounds on adversarial risk and showed that adversarial robustness is unattainable for any classifier with constant error, assuming the input data are sampled uniformly from n-spheres and the considered perturbation metric is Euclidean distance. Later on, [80] generalized their results for any concentrated metric probability space. Based on the assumption that the input data can be well captured by a smooth generative model, [44] proved that any classifier is vulnerable against adversarial examples with respect to the assumed input distribution. [108] showed that adversarial examples are inevitable, provided the maximum density of the input distribution is relatively small compared with uniform density. In this dissertation, I follow this line of works, but with a more practical goal to understand the intrinsic robustness for actual classification tasks of interest.

In addition to understanding the intrinsic robustness limits, another line of research studied a more comprehensive problem of adversarially robust generalization. For instance, [105] showed that compared with standard generalization, adversarially robust generalization requires significantly larger sample complexity for specific learning problems. To tackle such barrier, recent works [17, 1] demonstrated that by incorporating additional unlabeled data, semi-supervised learning methods can potentially bypass the sample complexity constraint. Other theoretical works [127, 65, 6] directly derive robust generalization bounds using extensions of Rademacher complexity under certain simplified settings, suggesting that compared with standard generalization, there may exist additional statistical barriers for adversarially robust generalization. In addition, [14] and [25] constructed specific tasks where computationally efficient robust classification is impossible, while [85] presented settings where adversarial robustness can only be achieved by improper learning. Other explanations for the difficulties of achieving adversarial robustness have been also proposed, such as the tension between standard and robust accuracy [119, 98], and the existence of well-generalizing but non-robust features [61], to name a few. However, all of the aforementioned theoretical analyses are conducted on special theoretical distributions, thus it is still unclear whether these results apply to actual datasets of interest.

Chapter 3

Intrinsic Robustness Limits

3.1 Introduction

The unsatisfactory adversarial robustness achieved by state-of-the-art machine learning classifiers motivates a fundamental information-theoretic question: *what are the inherent limitations of developing robust classifiers?* Several recent works [49, 44, 80, 108, 10] have shown that under certain assumptions regarding the data distribution and the perturbation metric, adversarial examples are theoretically inevitable. As a result, for a broad set of theoretically natural metric probability spaces of inputs, there is no classifier for the data distribution that achieves adversarial robustness. For example, [49] assumed that the input data are sampled uniformly from n-spheres and proved a model-independent theoretical bound connecting the risk to the average Euclidean distance to the "caps" (i.e., round regions on a sphere). [80] generalized this result to any concentrated metric probability space of inputs and showed, for example, that if the inputs come from any Normal Lévy family [74], any classifier with a noticable test error will be vulnerable to small perturbations.

Although such theoretical findings seem discouraging to the goal of developing robust classifiers, all these impossibility results depend on assumptions about data distributions that might not hold for cases of interest. In order to better understand the intrinsic limits of adversarial robustness for typical robust classification tasks of interest, we develop methods for testing properties of concrete datasets against these theoretical assumptions.

3.2 Preliminaries

Adversarial Risk. Adversarial risk captures the vulnerability of a classifier against adversarial perturbations. In particular, we adopt the following adversarial risk definition, which has been studied in several previous works including [49, 14, 80].

Definition 3.1 (Adversarial Risk)**.** Let $(\mathcal{X}, \mu)$ be the probability space of instances and $c : \mathcal{X} \to \mathcal{Y}$ be the ground-truth labeling function, where $\mathcal{Y}$ denotes the set of all possible labels. Consider perturbations with strength ϵ measured in distance metric Δ, then the *adversarial risk* of a classifier f is defined as:

$$\text{AdvRisk}_\epsilon(f; \mu, c, \Delta) = \Pr_{\boldsymbol{x} \sim \mu} \left[\exists\, \boldsymbol{x}' \in \text{Ball}(\boldsymbol{x}, \epsilon) \text{ s.t. } f(\boldsymbol{x}') \neq c(\boldsymbol{x}') \right].$$

When μ, c and Δ is free of context, we write $\text{AdvRisk}_\epsilon(f) = \text{AdvRisk}_\epsilon(f; \mu, c, \Delta)$ for simplicity. Correspondingly, the *adversarial robustness* of f is defined as:

$$\text{AdvRob}_\epsilon(f) = 1 - \text{AdvRisk}_\epsilon(f).$$

When $\epsilon = 0$, adversarial risk equals to the standard risk. Namely, $\text{AdvRisk}_0(f) = \text{Risk}(f) := \Pr_{\boldsymbol{x} \sim \mu}[f(\boldsymbol{x}) \neq c(\boldsymbol{x})]$ for any classifier f. Other definitions of adversarial risk have been proposed, such as the one used in [79]. These definitions are equivalent to the one we use, as long as small perturbations preserve the labels assigned by $c(\cdot)$[1].

Intrinsic Robustness. Given an adversarially robust classification problem, intrinsic robustness captures the maximum achievable adversarial robustness with respect to some family of classifiers.

[1] See [30] for a detailed comparison of these and other definitions of adversarial robustness.

Definition 3.2 (Intrinsic Robustness)**.** Consider the same setting as in Definition 3.1. Let $\mathcal{F}$ be some family of classifiers which map instances from $\mathcal{X}$ to $\mathcal{Y}$. The *intrinsic robustness* with respect to $\mathcal{F}$ is defined as:

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

According to the definition of intrinsic robustness, there does not exist any classifier in $\mathcal{F}$ with adversarial robustness higher than $\overline{\text{AdvRob}}_\epsilon(\mathcal{F})$ for the considered task.

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 3.3 (Concentration Function)**.** Consider a probability space $(\mathcal{X}, \mu)$ with distance metric Δ. For any $\epsilon > 0$ and $\alpha \in (0, 1)$, the *concentration function* is defined as:

$$h(\mu, \alpha, \epsilon; \Delta) = \inf_{\mathcal{E}\in\text{Pow}(\mathcal{X})} \{\mu(\mathcal{E}_\epsilon) \colon \mu(\mathcal{E}) \geq \alpha\}.$$

When the metric Δ is free of context, we write $h(\mu, \alpha, \epsilon) = h(\mu, \alpha, \epsilon; \Delta)$ for simplicity.

The standard notion of concentration function considers a special case of Definition 3.3 with $\alpha = 1/2$ (e.g., [117]). For some special metric probability spaces, one can prove the closed-form solution of the concentration function.

In particular, the Gaussian Isoperimetric Inequality [12, 114] characterizes the concentration function for spherical Gaussian distribution and ℓ_2-norm distance metric.

Lemma 3.4 (Gaussian Isoperimetric Inequality)**.** *Consider the standard Gaussian space $(\mathbb{R}^n, \nu_n)$ with ℓ_2-distance. Let $\mathcal{E} \in \text{Pow}(\mathbb{R}^n)$ and $\mathcal{H}$ be a half space such that $\nu_n(\mathcal{E}) =$*

$\nu_n(\mathcal{H})$*, then for any* $\epsilon \geq 0$*, it holds that*

$$\nu_n\big(\mathcal{E}_\epsilon^{(\ell_2)}\big) \geq \nu_n\big(\mathcal{H}_\epsilon^{(\ell_2)}\big) = \Phi\big(\Phi^{-1}\big(\nu_n(\mathcal{E})\big) + \epsilon\big).$$

Lemma 3.4 implies the closed-form solution of the concentration function with respect to the considered metric probability space. More formally, $h(\nu_n, \alpha, \epsilon; \ell_2) = \Phi(\Phi^{-1}(\alpha) + \epsilon)$.

3.2.1 Connecting Intrinsic Robustness with Concentration

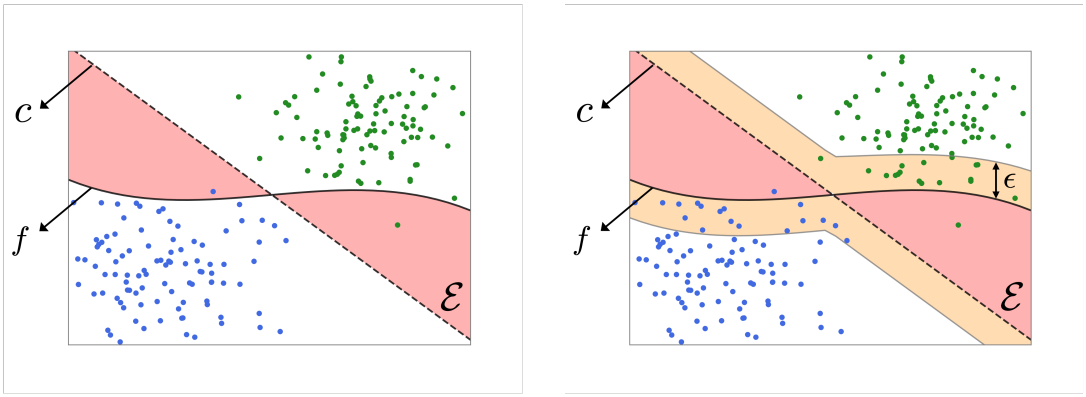
Let $(\mathcal{X}, \mu, \Delta)$ be the considered input metric probability space, $\mathcal{Y}$ be the set of possible labels, and $c : \mathcal{X} \to \mathcal{Y}$ be the concept function that gives each input a label. Given parameters $0 < \alpha < 1$ and $\epsilon \geq 0$, the concentration of measure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

$$\underset{\mathcal{E} \in \text{Pow}(\mathcal{X})}{\text{minimize}} \ \mu(\mathcal{E}_\epsilon) \quad \text{subject to} \quad \mu(\mathcal{E}) \geq \alpha. \tag{3.1}$$

For any classifier f, let $\mathcal{E} = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \neq c(\boldsymbol{x})\}$ be its induced error region with respect to $c(\cdot)$. By connecting the risk of f with the measure of $\mathcal{E}$ and the adversarial risk of f with the measure of the ϵ-expansion of $\mathcal{E}$ (see Figure 3.1 for the illustration of these connections), [80] proved that the concentration of measure problem is equivalent to the following optimization problem regarding risk and adversarial risk:

$$\underset{f}{\text{minimize}} \ \text{AdvRisk}_\epsilon(f) \quad \text{subject to} \quad \text{Risk}(f) \geq \alpha. \tag{3.2}$$

More specifically, the following theorem, proven in [80], characterizes the fundamental connection between the concentration of measure and an intrinsic robustness limit.



(a) $\text{Risk}(f) = \mu(\mathcal{E})$ (b) $\text{AdvRisk}(f) = \mu\left(\mathcal{E}_\epsilon^{(\Delta)}\right)$

Figure 3.1: Illustration of the connection between risk and error region as well as the connection between adversarial risk and expanded error region: (a) the risk of any classifier f is equivalent to the measure of its induced error region $\mathcal{E}$, (b) the adversarial risk of any classifier f corresponds to the measure of the ϵ-expansion of its error region $\mathcal{E}_\epsilon^{(\Delta)}$. Here Δ is set as the ℓ_2-norm distance metric in the Figure 3.1(b).

Theorem 3.5. *Consider input metric probability space $(\mathcal{X}, \mu, \Delta)$, label space $\mathcal{Y}$ and ground-truth labeling function c. For any classifier $f : \mathcal{X} \to \mathcal{Y}$ and $\epsilon \geq 0$, it holds that*

$$\text{AdvRisk}_\epsilon(f) \geq h(\mu, \text{Risk}(f), \epsilon).$$

For any $\alpha \in (0, 1)$, if denote by $\mathcal{F}_\alpha = \{f : \text{Risk}(f) \geq \alpha\}$ the set of classifiers with imperfect risk, then it holds for any $\epsilon \geq 0$ that

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

Theorem 3.5 suggests that the concentration function of the input metric probability space can be translated into an adversarial robustness upper bound that applies to any classifier with risk at least α. If this upper bound is small, then one can conclude that it is impossible to learn an adversarially robust classifier, as long as the classifier has risk at least α.

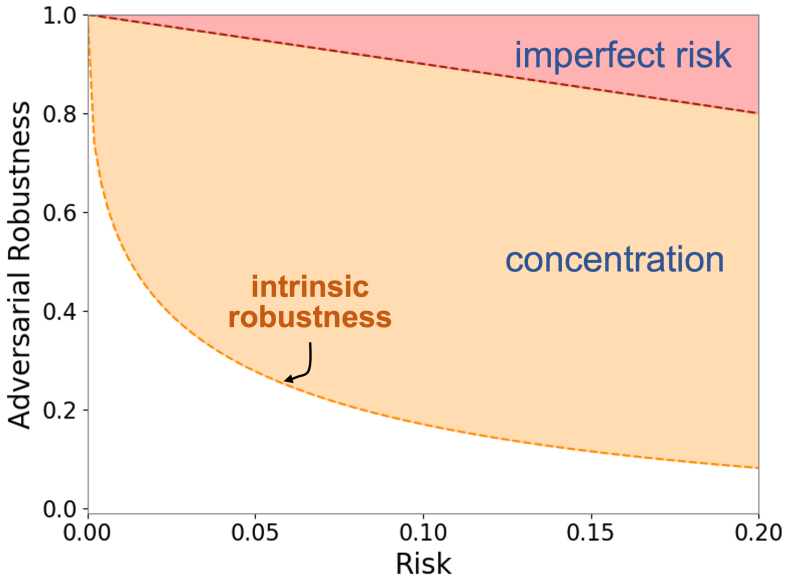


Figure 3.2: Illustration of the theoretical upper bounds on adversarial robustness for the robust classification task on uniform n-spheres considered in [49]. The red dashed line represents the naive upper bound, namely $\text{AdvRob}_\epsilon(f) \leq 1 - \text{Risk}(f)$ for any classifier f; whereas the orange dashed line denotes the intrinsic robustness limit translated by concentration of measure, namely $\text{AdvRob}_\epsilon(f) \leq 1 - h(\mu, \text{Risk}(f), \epsilon)$ for any classifier f.

To give a specific example, we visualize the intrinsic robustness limit translated by concentration function for the synthetic robust classification problem studied in [49] in Figure 3.2. In particular, the inputs are assumed to be uniformly distributed over two concentric n-spheres and ℓ_2-norm bounded perturbations are considered[2]. Here, we set $n = 1000$ and consider the ℓ_2 perturbations with strength $\epsilon = 0.1$ for illustration. Figure 3.2 suggests that due to the concentration of measure phenomenon, classifiers are only attainable with risk and adversarial robustness below the orange dashed line for the considered robust classification task. In particular, as long as a classifier has risk larger than 1%, the maximum achievable adversarial robustness implied by the concentration of measure phenomenon is less than 54%, suggesting that adversarial robustness is inherently unattainable for classifying inputs generated according to the considered setting.

[2]See [49] for the detailed description of the considered classification problem

3.3 Conditional Generative Model based Approach[3]

Although closed-form solutions to the concentration of measure problem (3.1) can be proved with respect to some specific metric probability spaces, it remains unclear about the implications of Theorem 3.5 on intrinsic robustness for general distributions such as images. In this section, we leverage the power of conditional generative models for modeling natural image distributions, which builds a bridge to understand the intrinsic robustness limits of general distributions using the concentration property of well-behaved latent space.

3.3.1 Definitions and Assumptions

Conditional Generative Models. Motivated by the great success of producing natural-looking images using conditional generative adversarial nets (GANs) [83, 92, 13], we assume that the underlying data distribution μ can be modeled by some *conditional generative model*. A generative model can be seen as a function $g : \mathcal{Z} \to \mathcal{X}$ that maps some latent distribution, usually assumed to be multivariate Gaussian, to some generated distribution.

Conditional generative models incorporate the additional class information into the data generating process. A conditional generative model can be considered as a set of generative models $\{g_i\}_{i\in[K]}$, where images from the i-th class can be generated by transforming latent Gaussian vectors through g_i. More rigorously, we say a probability distribution μ can be generated by a conditional generative model $\{(g_i, p_i)\}_{i\in[K]}$, if $\mu = \sum_{i=1}^{K} p_i \cdot (g_i)_*(\nu_d)$, where K is the total number of different class labels, and $p_i \in [0, 1]$ represents the probability of

[3]Xiao Zhang*, Jinghui Chen*, Quanquan Gu, David Evans, *Understanding the Intrinsic Robustness of Image Distributions using Conditional Generative Models*, in the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020) [131].

sampling an image from class i.

In-distribution Adversarial Risk. Given the conditional generative process, *in-distribution adversarial risk* captures the vulnerability of a classifier against adversarial examples that lie on the generated image manifold.

Definition 3.6 (In-distribution Adversarial Risk)**.** Consider the same settings as in Definition 3.1. Suppose μ can be captured by a conditional generative model $\{(g_i, p_i)\}_{i\in[K]}$. For any given classifier f, the *in-distribution adversarial risk* of f against ϵ-perturbations measured by metric Δ is defined as:

$$\text{In-AdvRisk}_\epsilon(f) = \Pr_{(\boldsymbol{x},i)\sim\mu}\left[\exists\, \boldsymbol{z}' \in \mathcal{Z} \text{ s.t. } g_i(\boldsymbol{z}') \in \mathcal{B}_\epsilon(\boldsymbol{x}) \text{ and } f(g_i(\boldsymbol{z}')) \neq c(g_i(\boldsymbol{z}'))\right].$$

Given that the in-distribution adversarial risk restricts the adversarial examples to be on the image manifold, it holds that, for any classifier f, $\text{In-AdvRisk}_\epsilon(f) \leq \text{AdvRisk}_\epsilon(f)$.

Local Lipschitz Condition. The *local Lipschitz condition* characterizes the smoothness property of a generative model, which connects perturbations in the image space to perturbations in the latent space.

Condition 3.3.1 (Local Lipschitz Condition)**.** Let $g : \mathbb{R}^d \to \mathcal{X}$ be a generative model that maps the latent Gaussian distribution ν_d to some generated distribution. Consider Euclidean distance as the distance metric for $\mathbb{R}^d$, and Δ as the metric for $\mathcal{X}$. Given $r > 0$ and $0 < \delta < 1$, g is said to be $L(r)$-locally Lipschitz with probability at least $1 - \delta$, if it satisfies

$$\Pr_{\boldsymbol{z}\sim\nu_d}\left[\forall \boldsymbol{z}' \in \mathcal{B}_r(\boldsymbol{z}),\ \Delta\big(g(\boldsymbol{z}'), g(\boldsymbol{z})\big) \leq L(r) \cdot \|\boldsymbol{z}' - \boldsymbol{z}\|_2\right] \geq 1 - \delta.$$

3.3.2 Theoretical Results on Intrinsic Robustness

Making use of the Gaussian Isoperimetric Inequality (Lemma 3.4) and local Lipschitz property of the conditional generator (Condition 3.3.1), the following theorem proves a lower bound on the (in-distribution) adversarial risk for any given classifier, provided that the underlying distribution can be captured by a conditional generative model.

Theorem 3.7. *Let $(\mathcal{X}, \mu, \Delta)$ be a metric probability space and $c : \mathcal{X} \to [K]$ be the underlying ground-truth. Suppose μ can be generated by a conditional generative model $\{(g_i, p_i)\}_{i\in[K]}$. Given $\epsilon > 0$, suppose there exist constants $r > 0$ and $\delta \in (0, 1)$ such that for any $i \in [K]$, g_i satisfies $L_i(r)$-local Lipschitz property with probability at least $1 - \delta$ and $r \cdot L_i(r) \geq \epsilon$. Then for any classifier f, it holds that*

$$\text{AdvRisk}_\epsilon(f) \geq \textit{In-AdvRisk}_\epsilon(f) \geq \sum_{i=1}^{K} p_i \cdot \Phi\left(\Phi^{-1}\big(\text{Risk}(f; \mu_i)\big) + \frac{\epsilon}{L_i(r)}\right) - \delta,$$

where $\mu_i = (g_i)_*(\nu_d)$ is the push-forward measure of ν_d though g_i, for any $i \in [K]$.

Proof of Theorem 3.7. Let $\mathcal{E} = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \neq c(\boldsymbol{x})\}$ be the error region in the image space and $\mathcal{E}_\epsilon$ be the ϵ-expansion of $\mathcal{E}$ with metric Δ. By Definition 3.1, we have

$$\text{AdvRisk}_\epsilon(f; \mu) = \mu(\mathcal{E}_\epsilon) = \sum_{i=1}^{K} p_i \cdot \mu_i(\mathcal{E}_\epsilon) = \sum_{i=1}^{K} p_i \cdot \text{AdvRisk}_\epsilon(f; \mu_i).$$

According to Definition 3.6, we have $\text{AdvRisk}_\epsilon(f; \mu_i) \geq \text{In-AdvRisk}_\epsilon(f; \mu_i)$ for any $i \in [K]$. Thus, it remains to lower bound each term $\text{In-AdvRisk}_\epsilon(f; \mu_i)$ individually. For any

classifier f, we have

$$\begin{aligned}\text{In-AdvRisk}_\epsilon(f;\mu_i) &= \Pr_{\boldsymbol{z}\sim\nu_d}\Big[\exists\, \boldsymbol{z}' \in \mathbb{R}^d,\ \text{s.t. } \Delta\big(g_i(\boldsymbol{z}'), g_i(\boldsymbol{z})\big) \le \epsilon, f\big(g_i(\boldsymbol{z}')\big) \neq c\big(g_i(\boldsymbol{z}')\big)\Big] \\ &\ge \underbrace{\Pr_{\boldsymbol{z}\sim\nu_d}\Big[\exists\, \boldsymbol{z}' \in \mathcal{B}_{\epsilon/L_i(r)}(\boldsymbol{z}),\ \text{s.t. } f\big(g_i(\boldsymbol{z}')\big) \neq c\big(g_i(\boldsymbol{z}')\big)\Big]}_{I} - \delta \qquad (3.3)\end{aligned}$$

where the first inequality is due to $\mu_i = (g_i)_*(\nu_d)$, and the second inequality holds because g_i satisfies the $L_i(r)$-locally Lipschitz condition and $\mathcal{B}_{\epsilon/L_i(r)}(\boldsymbol{z}) \subseteq \mathcal{B}_r(\boldsymbol{z})$ holds for any $\boldsymbol{z}$.

To further bound the term I, we make use of the Gaussian Isoperimetric Inequality as presented in Lemma 3.4. Let $\mathcal{A}_f = \{\boldsymbol{z} \in \mathbb{R}^d : f(g_i(\boldsymbol{z})) \neq c(g_i(\boldsymbol{z}))\}$ be the corresponding error region in the latent space. By Lemma 3.4, we have

$$I \ge \Phi\bigg(\Phi^{-1}\big(\nu_d(\mathcal{A}_f)\big) + \frac{\epsilon}{L_i(r)}\bigg) = \Phi\bigg(\Phi^{-1}\big(\text{Risk}(f;\mu_i)\big) + \frac{\epsilon}{L_i(r)}\bigg). \qquad (3.4)$$

Finally, plugging (3.4) into (3.3), we complete the proof. □

Theorem 3.7 suggests the (in-distribution) adversarial risk is related to the risk on each data manifold and the ratio between the perturbation strength and the local Lipschitz constant.

The following theorem gives a theoretical upper bound on the intrinsic robustness with respect to the family of classifiers with imperfect risk.

Theorem 3.8. *Under the same setting as in Theorem 3.7, let* $L_{\max}(r) = \max_{i\in[K]} L_i(r)$. *Consider the class of imperfect classifiers* $\mathcal{F}_\alpha = \{f : \text{Risk}(f) \ge \alpha\}$ *with* $\alpha > 0$, *then the intrinsic robustness with respect to* $\mathcal{F}_\alpha$ *can be bounded as,*

$$\overline{\text{AdvRob}}_\epsilon(\mathcal{F}_\alpha) \le 1 + \delta - \min_{i\in[K]}\left\{p_i \cdot \Phi\left(\Phi^{-1}\left(\frac{\alpha}{p_i}\right) + \frac{\epsilon}{L_{\max}(r)}\right)\right\},$$

provided that $\alpha/p_i \leq 1$ *for any* $i \in [K]$. *In addition, if we consider the family of classifiers that have conditional risk at least* α *for each class, namely* $\tilde{\mathcal{F}}_\alpha = \{f : \text{Risk}(f; \mu_i) \geq \alpha, \forall i \in [K]\}$, *then the intrinsic robustness with respect to* $\tilde{\mathcal{F}}_\alpha$ *can be bounded by*

$$\overline{\text{AdvRob}}_\epsilon(\tilde{\mathcal{F}}_\alpha) \leq 1 + \delta - \sum_{i=1}^{K} p_i \cdot \Phi\bigg(\Phi^{-1}\big(\alpha\big) + \frac{\epsilon}{L_{\max}(r)}\bigg).$$

Proof of Theorem 3.8. According to Theorem 3.7, for any $f \in \mathcal{F}_\alpha$, we have

$$\begin{aligned}\overline{\text{AdvRob}}_\epsilon(\mathcal{F}_\alpha) &\leq 1 + \delta - \sum_{i=1}^{K} p_i \cdot \Phi\bigg(\Phi^{-1}\big(\text{Risk}(f; \mu_i)\big) + \frac{\epsilon}{L_i(r)}\bigg) \\ &\leq 1 + \delta - \sum_{i=1}^{K} p_i \cdot \Phi\bigg(\Phi^{-1}\big(\text{Risk}(f; \mu_i))\big) + \frac{\epsilon}{L_{\max}(r)}\bigg), \end{aligned} \tag{3.5}$$

where the last inequality holds because $\Phi(\cdot)$ is monotonically increasing. For any $f \in \mathcal{F}_\alpha$, let $\mathcal{E} = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \neq c(\boldsymbol{x})\}$ be the error region and $\alpha_i = \mu_i(\mathcal{E})$ be the measure of $\mathcal{E}$ under the i-th conditional distribution.

Thus, to obtain an upper bound on $\overline{\text{AdvRob}}_\epsilon(\mathcal{F}_\alpha)$ using (3.5), it remains to solve the following optimization problem:

$$\underset{\alpha_1,\ldots,\alpha_K \in [0,1]}{\text{minimize}} \sum_{i=1}^{K} p_i \cdot \Phi\bigg(\Phi^{-1}(\alpha_i) + \frac{\epsilon}{L_{\max}(r)}\bigg) \quad \text{subject to } \sum_{i=1}^{K} p_i \alpha_i \geq \alpha. \tag{3.6}$$

Note that for classifier in $\tilde{\mathcal{F}}_\alpha$, by definition, we can simply replace $\alpha_i = \alpha$ in (3.6), which proves the upper bound on $\overline{\text{AdvRob}}_\epsilon(\tilde{\mathcal{F}}_\alpha)$.

Next, we are going to show that the optimal value of (3.6) is achieved, only if there exists a class $i' \in [K]$ such that $\alpha_{i'} = \alpha/p_{i'}$ and $\alpha_i = 0$ for any $i \neq i'$. Consider the simplest case where $K = 2$. Note that $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are both monotonically increasing functions, which implies that $\sum_{i=1}^{K} p_i \alpha_i = \alpha$ holds when optimum achieved, thus the optimization

problem for $K = 2$ can be formulated as follows

$$\min_{\alpha_1,\alpha_2\in[0,1]} \sum_{i\in\{1,2\}} p_i \cdot \Phi\left(\Phi^{-1}(\alpha_i) + \frac{\epsilon}{L_{\max}(r)}\right) \quad \text{s.t.} \sum_{i\in\{1,2\}} p_i\alpha_i = \alpha. \tag{3.7}$$

Suppose $\alpha_1 \geq \alpha_2$ holds for the initial setting. Now consider another setting where $\alpha_1' > \alpha_1$, $\alpha_2' < \alpha_2$. Let $s_1 = \Phi^{-1}(\alpha_1') - \Phi^{-1}(\alpha_1)$ and $s_2 = \Phi^{-1}(\alpha_2) - \Phi^{-1}(\alpha_2')$. According to the equality constraint of the optimization problem (3.7), we have

$$p_1 \cdot \int_{\Phi^{-1}(\alpha_1)}^{\Phi^{-1}(\alpha_1)+s_1} \frac{1}{\sqrt{2\pi}} \cdot \exp^{-x^2/2} dx = p_2 \cdot \int_{\Phi^{-1}(\alpha_2)-s_2}^{\Phi^{-1}(\alpha_2)} \frac{1}{\sqrt{2\pi}} \cdot \exp^{-x^2/2} dx. \tag{3.8}$$

Let $\eta = \epsilon / L_{\max}(r)$ for simplicity. By simple algebra, we have

$$\begin{aligned}
p_1 \cdot \int_{\Phi^{-1}(\alpha_1)+\eta}^{\Phi^{-1}(\alpha_1)+s_1+\eta} \frac{1}{\sqrt{2\pi}} \cdot \exp^{-x^2/2} dx &= p_1 \cdot \int_{\Phi^{-1}(\alpha_1)}^{\Phi^{-1}(\alpha_1)+s_1} \frac{1}{\sqrt{2\pi}} \cdot \exp^{-u^2/2-\eta\cdot u-\eta^2/2} du \\
&< p_1 \cdot \exp^{-\eta\cdot\Phi^{-1}(\alpha_1)-\eta^2/2} \cdot \int_{\Phi^{-1}(\alpha_1)}^{\Phi^{-1}(\alpha_1)+s_1} \frac{1}{\sqrt{2\pi}} \cdot \exp^{-u^2/2} du \\
&\leq p_2 \cdot \exp^{-\eta\cdot\Phi^{-1}(\alpha_2)-\eta^2/2} \cdot \int_{\Phi^{-1}(\alpha_2)-s_2}^{\Phi^{-1}(\alpha_2)} \frac{1}{\sqrt{2\pi}} \cdot \exp^{-u^2/2} du \\
&< p_2 \cdot \int_{\Phi^{-1}(\alpha_2)-s_2+\eta}^{\Phi^{-1}(\alpha_2)+\eta} \frac{1}{\sqrt{2\pi}} \cdot \exp^{-x^2/2} dx,
\end{aligned}$$

where the first inequality holds because $\exp^{-\eta\cdot u} < \exp^{-\eta\cdot\Phi^{-1}(\alpha_1)}$ for any $u > \Phi^{-1}(\alpha_1)$, the second inequality follows from (3.8) and the fact that $\Phi^{-1}(\alpha_1) \geq \Phi^{-1}(\alpha_2)$, and the last inequality holds because $\exp^{-\eta\cdot\Phi^{-1}(\alpha_2)} < \exp^{-\eta\cdot u}$ for any $u < \Phi^{-1}(\alpha_2)$. Therefore, the optimal value of (3.7) will be achieved when $\alpha_1 = 0$ or $\alpha_2 = 0$. For general setting with $K > 2$, since $\alpha_1, \ldots, \alpha_K$ are independent in the objective, we can fix $\alpha_3, \ldots, \alpha_K$ and optimize α_1 and α_2 first, then deal with α_i incrementally using the same technique. □

Remark 3.9. Theorem 3.8 shows that if the data distribution can be captured by a condi-

tional generative model, the intrinsic robustness bound with respect to imperfect classifiers will largely depend on the ratio ϵ/L_{max}. For instance, if we assume the ratio $\epsilon/L_{\text{max}} = 1$, then Theorem 3.8 suggests that no classifier with initial risk at least 5% can achieve robust accuracy exceeding 75%. In addition, if we assume the local Lipschitz parameter L_{max} is some constant, then adversarial robustness is indeed not achievable for high-dimensional data distributions, provided the perturbation strength ϵ is sublinear to the input dimension, which is the typical setting considered in previous works [49, 44, 80].

Remark 3.10. The intrinsic robustness is closely related to the in-distribution adversarial risk. For the class of classifiers $\mathcal{F}_\alpha$, one can prove that the intrinsic robustness is equivalent to the maximum achievable in-distribution adversarial robustness:

$$\overline{\text{AdvRob}}_\epsilon(\mathcal{F}_\alpha) = 1 - \inf_{f \in \mathcal{F}_\alpha} \{\text{In-AdvRisk}_\epsilon(f)\}. \tag{3.9}$$

On one hand, $\text{AdvRisk}_\epsilon(f) \geq \text{In-AdvRisk}_\epsilon(f)$ holds for any f. On the other hand, for any $f \in \mathcal{F}_\alpha$, one can construct an $h_f \in \mathcal{F}_\alpha$ such that $h_f(\boldsymbol{x}) = f(\boldsymbol{x})$ if $\boldsymbol{x} \in \mathcal{E}_f \cap \mathcal{M}$ and $h_f(\boldsymbol{x}) = c(\boldsymbol{x})$ otherwise, where $\mathcal{E}_f = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \neq c(\boldsymbol{x})\}$ denotes the error region of f and $\mathcal{M}$ is the considered image manifold. The construction immediately suggests $\text{In-AdvRisk}_\epsilon(f) = \text{AdvRisk}_\epsilon(h_f)$, which implies,

$$\inf_{f \in \mathcal{F}_\alpha} \{\text{In-AdvRisk}_\epsilon(f)\} = \inf_{f \in \mathcal{F}_\alpha} \{\text{AdvRisk}_\epsilon(h_f)\} \geq \inf_{f \in \mathcal{F}_\alpha} \{\text{AdvRisk}_\epsilon(f)\}.$$

Combining both directions proves the soundness of (3.9). This equivalence suggests the in-distribution adversarial robustness of any classifier in $\mathcal{F}_\alpha$ can be viewed as a lower bound on the actual intrinsic robustness, which motivates us to study the intrinsic robustness by estimating the in-distribution adversarial robustness of trained models in our experiments.

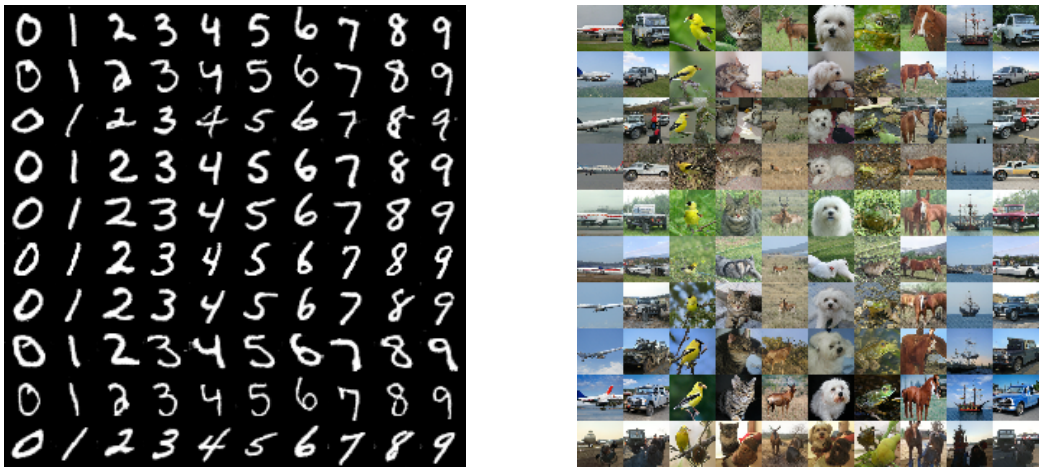
3.3.3 Experiments

This section provides our empirical evaluations of the intrinsic robustness on typical image distributions to evaluate the tightness of our bound. We test our bound on two image distributions generated using MNIST [72] and ImageNet [26] datasets.

Conditional GAN Models. Instead of directly evaluating the robustness on real datasets, we make use of conditional GAN models to generate datasets from the learned data distributions and evaluate the robustness of several state-of-the-art robust models trained on the generated dataset for a fair comparison with the theoretical robustness limits. Note that this approach is only feasible with conditional generative models as unconditional models cannot provide the corresponding labels for the generated data samples. For MNIST, we adopt ACGAN [92] which features an additional auxiliary classifier for better conditional image generation. The ACGAN model generates 28×28 images from a 100-dimension latent space concatenated with an addition 10-dimension one-hot encoding of the conditional class labels. For ImageNet, we adopt the BigGAN model [13] which is the state-of-the-art GAN model in conditional image generation. It generates 128×128 images from a 120-dimension latent space. We down-sampled the generated images to 32×32 for efficiency propose. We consider a standard Gaussian[4] as the latent distribution for both conditional generative models. Figure 3.3 shows examples of the generated MNIST and ImageNet images. For both figures, each column of images corresponds to a particular label class of the considered dataset.

Local Lipschitz Constant Estimation. From Theorem 3.8, we observe that given a class of classifiers with risk at least α, the derived intrinsic robustness upper bound is mainly decided by the perturbation strength ϵ and the local Lipschitz constant $L(r)$. While ϵ is usually

[4]The original BigGAN model uses truncated Gaussian. We adapted it to standard Gaussian distribution.



(a) ACGAN Generated MNIST (b) BigGAN Generated ImageNet

Figure 3.3: Illustration of the generated images using different conditional models. For BigGAN generated images, we select 10 specific classes from the 1000 ImageNet classes (corresponding to the 10 image classes in CIFAR-10).

predesignated in common robustness evaluation settings, the local Lipschitz constant $L(r)$ is unknown for most real world tasks. Computing an exact Lipschitz constant of a deep neural network is a difficult open problem. Thus, instead of obtaining the exact value, we approximate $L(r)$ using a sample-based approach with respect to the generative models.

Recalling Definition 3.3.1, we consider Δ as the ℓ_2 distance and $g(\boldsymbol{z})$ and $g(\boldsymbol{z}')$ are easy to compute via the generator network. Computing $L(r)$, however, is much more complicated as it requires obtaining a maximum value within a radius-r ball. To deal with this, our approach approximates $L(r)$ by sampling N points in the neighborhood around $\boldsymbol{z}$ and takes the maximum value as the estimation of the true maximum value within the ball. Since the definition of local Lipschitz is probabilistic, we take multiple samples of the latent vectors $\boldsymbol{z}$ to estimate the local Lipschitz constant $L(r)$. The estimation procedure is summarized in Algorithm 1, which gives an underestimate of the underlying truth. Developing better Lipschitz estimation methods is an active area in machine learning research, but is not the main focus of this work.

Algorithm 1: Local Lipschitz Estimation

Input : number of samples S, number of local neighbors per sample N, r, δ

for $i = 1, \ldots, S$ **do**

  Generate a latent space sample $\boldsymbol{z}_i$;

  Generate N samples $\{\hat{\boldsymbol{z}}_i^j\}_{j=1}^N$ within $\mathcal{B}_r(\boldsymbol{z}_i)$;

  $L_i = \max_j \frac{\|g(\hat{\boldsymbol{z}}_i^j) - g(\boldsymbol{z}_i)\|_2}{\|\hat{\boldsymbol{z}}_i^j - \boldsymbol{z}_i\|_2}$;

end

Output : $(1-\delta)$-percentile of $\{L_i\}_{i=1}^S$

Tables 3.1 and 3.2 summarize the local Lipschitz constants estimated for the trained ACGAN and BigGAN generators conditioned on each class. In particular, we report both the mean estimates averaged over 10 repeated trials and the standard deviations. For both conditional generators, we set $S = 1000$, $N = 2000$, $r = 0.5$ and $\delta = 0.001$ in Algorithm 1 for Lipschitz estimation. For BigGAN, the specifically selected 10 classes from ImageNet are reported in Table 3.2.

Compared with unconditional generative models, conditional ones generate each class using a separate generator. Thus, the local Lipschitz constant of each class-conditioned generator is expected to be smaller than that of unconditional ones, as the within-class variation is usually much smaller than the between-class variation for a given classification dataset. For instance, we trained an unconditional GAN generator [50] on MNIST dataset, which yields an overall local Lipschitz constant of 27.01 from Algorithm 1 under the same parameter settings. If we plug in this estimated Lipschitz constant into the theoretical results in [44], the implied intrinsic robustness bound is in fact vacuous (above 1) with perturbations strength $\epsilon \leq 3.0$ in ℓ_2 distance.

Comparisons with Robust Classifiers. We compare our derived intrinsic robustness upper bound with the empirical adversarial robustness achieved by the current state-of-the-art defense methods under ℓ_2 perturbations. Specifically, we consider three robust training

Table 3.1: The estimated local Lipschitz constants of the trained ACGAN model on the 10 MNIST classes with $r = 0.5$ and $\delta = 0.001$.

Class	digit 0	digit 1	digit 2	digit 3	digit 4
Lipschitz	7.9 ± 0.3	8.6 ± 0.4	8.3 ± 0.4	7.8 ± 0.3	10.3 ± 0.6
Class	digit 5	digit 6	digit 7	digit 8	digit 9
Lipschitz	11.0 ± 0.4	9.5 ± 0.3	7.8 ± 0.2	9.3 ± 0.4	10.9 ± 0.4

Table 3.2: The estimated local Lipschitz constants of the BigGAN model on the 10 selected ImageNet classes with $r = 0.5$ and $\delta = 0.001$.

Class	airliner	jeep	goldfinch	tabby cat	hartebeest
Lipschitz	13.1 ± 0.8	14.5 ± 1.1	11.7 ± 0.5	12.4 ± 0.4	10.4 ± 1.1
Class	Maltese dog	bullfrog	sorrel	pirate ship	pickup
Lipschitz	11.3 ± 0.6	9.4 ± 0.3	13.0 ± 0.3	13.1 ± 0.8	14.9 ± 0.9

methods: *LP-Certify*: optimization-based certified robust defense [122]; *Adv-Train*: PGD attack based adversarial training [79]; and *TRADES*: adversarial training by accuracy and robustness trade-off [129]. We adopt these robust training methods to train robust classifiers over a set of generated training images and evaluate their robustness on the corresponding generated test set.

For MNIST, we use our trained ACGAN model to generate 10 classes of hand-written digits with $60,000$ training images and $10,000$ testing images. For ImageNet, we use the BigGAN model to generate 10 selected classes of images, which contains $50,000$ images for training set and $10,000$ images for test set. We refer to the 10-class BigGAN generated dataset as 'ImageNet10'. We set $\epsilon = 3.0$ for training robust models using Adv-Train and TRADES for both generated datasets, whereas we only train the LP-based certified robust classifier with $\epsilon = 2.0$ on generated MNIST data, as it is not able to scale with ImageNet10 as well as generated MNIST with larger ϵ.

Table 3.3: Comparisons between the empirically measured robustness of adversarially trained classifiers and the implied theoretical intrinsic robustness bound on the conditional generated datasets.

Dataset	Method	Natural Accuracy	Adversarial Robustness		
			$\epsilon = 1.0$	$\epsilon = 2.0$	$\epsilon = 3.0$
Generated MNIST	LP-Certify	$88.3 \pm 0.2\%$	$74.0 \pm 0.4\%$	$51.1 \pm 0.6\%$	$23.5 \pm 0.3\%$
	Adv-Train	$97.2 \pm 0.2\%$	$93.1 \pm 0.2\%$	$83.5 \pm 0.3\%$	$58.9 \pm 0.4\%$
	TRADES	$98.3 \pm 0.1\%$	$94.8 \pm 0.2\%$	$81.8 \pm 0.4\%$	$57.7 \pm 0.4\%$
	Our Bound	-	98.2%	97.8%	97.2%
ImageNet10	Adv-Train	$82.1 \pm 0.3\%$	$67.8 \pm 0.3\%$	$47.1 \pm 0.4\%$	$23.4 \pm 0.4\%$
	TRADES	$83.4 \pm 0.3\%$	$68.5 \pm 0.3\%$	$49.1 \pm 0.5\%$	$27.8 \pm 0.5\%$
	Our Bound	-	83.5%	81.8%	80.0%

A commonly-used method to evaluate the robustness of a given model is by performing carefully-designed adversarial attacks. Here we adopt the PGD attack [79], and report the robust accuracy (classification accuracy on inputs generated using the PGD attack) as the empirically measured model robustness. We test both the natural classification accuracy and the robustness of the aforementioned adversarially trained classifiers under ℓ_2 perturbations with perturbation strength ϵ selected from $\{1.0, 2.0, 3.0\}$.

Table 3.3 compares the empirically measured robustness of the trained robust classifiers and the derived theoretical upper bound on intrinsic robustness. For empirically measured adversarial robustness, we report both the mean and the standard deviation with respect to 10 repeated trials. For computing our theoretical robust bounds, we plug the estimated local Lipschitz constants into Theorem 3.8 with risk threshold $\alpha = 0.015$ for generated MNIST and $\alpha = 0.15$ for ImageNet10, to reflect the best natural accuracy achieved by the considered robust classifiers.

Under most settings, there exists a large gap between the robust limit implied by our theory and the best adversarial robustness achieved by state-of-the-art robust classifiers. For

instance, Adv-Train and TRADES only achieve less than 50% robust accuracy on the generated ImageNet10 data with $\epsilon = 2.0$, whereas the estimated robustness bound is as high as 81.8%. The gap becomes even larger when we increase the perturbation strength ϵ. In contrast to the previous theoretical results on artificial distributions, for these image classification problems we cannot simply conclude from the intrinsic robustness bound that adversarial examples are inevitable. This huge gap between the empirical robustness of the best current image classifiers and the estimated theoretical bound suggests that either there is a way to train better robust models or that there exist other explanations for the inherent limitations of robust learning against adversarial examples.

In-distribution Adversarial Robustness. In previous sections, we empirically show the unconstrained robustness of existing robust classifiers is far below the intrinsic robustness upper bound implied by our theory for real distributions. However, it is not clear whether the reason is that current robust training methods are far from perfect, or that our derived upper bound is not tight enough due to the Lipschitz relaxation step used for proving such bound. In this section, we empirically study the in-distribution adversarial risk for a better characterization of the actual intrinsic robustness. As shown in Remark 3.10, the in-distribution adversarial robustness of any classifier with risk at least α can be regarded as a lower bound for the intrinsic robustness $\overline{\text{AdvRob}}_\epsilon(\mathcal{F}_\alpha)$. This provides us a more accurate characterization of the intrinsic robustness bound and enables better understanding of intrinsic robustness.

While there are many types of attack algorithms in the literature that can be used to evaluate the unconstrained robustness of a given classifier in the image space, little has been done in terms of how to evaluate the in-distribution robustness. In order to empirically evaluate the in-distribution robustness, we straightforwardly formulate the following optimization

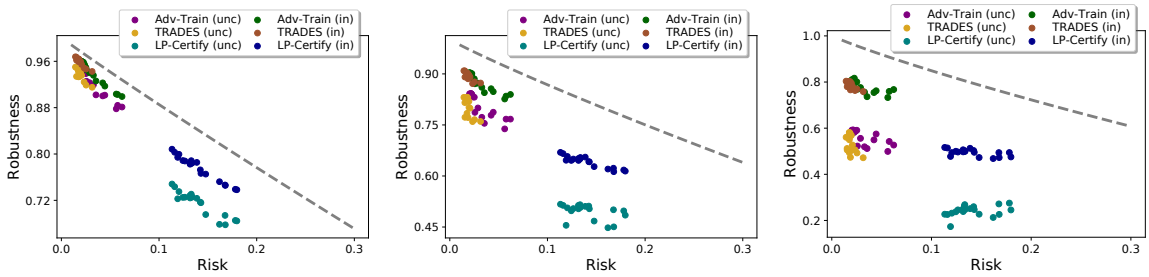
problem to find adversarial examples on the image manifold:

$$\min_{\boldsymbol{z}} \ \mathcal{L}(f(G(\boldsymbol{z}, y)), y) \quad \text{s.t.} \quad \|G(\boldsymbol{z}, y) - \boldsymbol{x}\|_2 \leq \epsilon, \tag{3.10}$$

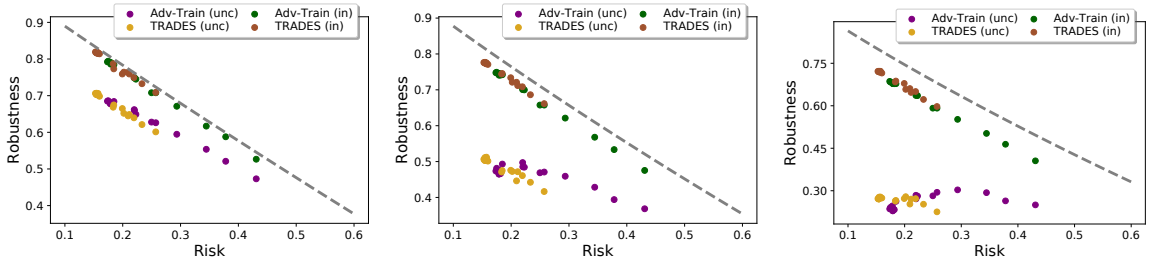
where $\boldsymbol{z} \in \mathbb{R}^d$, $\boldsymbol{x}$ is the data sample in the image space to be attacked, f is the given classifier, and $\mathcal{L}$ denotes the adversarial loss function. The goal of (3.10) is to optimize the latent vector to lower the adversarial loss (make the robust classifier mis-classify some generated images) while keeping the distance between the generated image and the test image within ϵ perturbation limit. The key difficulty in solving (3.10) lies in the fact that we cannot perform any type of projection operations as we are optimizing over $\boldsymbol{z}$ but the constraints are imposed on the generated image space $G(\boldsymbol{z}, y)$. This prohibits the use of common attack algorithms such as PGD. In order to solve (3.10), we transform (3.10) into the following Lagrangian formulation:

$$\min_{\boldsymbol{z}} \|G(\boldsymbol{z}, y) - \boldsymbol{x}\|_2 + \lambda \cdot \mathcal{L}(f(G(\boldsymbol{z}, y)), y). \tag{3.11}$$

This formulation ignores the perturbation constraint of ϵ and tries to find the in-distribution adversarial examples with the smallest possible perturbation. In order to evaluate the intrinsic robustness under a given ϵ perturbation budget, we need to further check all in-distribution adversarial examples found and only count those with perturbations within the ϵ constraint. Note that even though (3.11) provides us a feasible way to compute the in-distribution robustness of a classifier, equation (3.11) itself could be hard to solve in general. First, it is not obvious how to initialize $\boldsymbol{z}$. Random initialization of $\boldsymbol{z}$ could lead to bad local optima which prevent the optimizer from efficiently solving (3.11) or even finding a $\boldsymbol{z}$ that could make $G(\boldsymbol{z}, y)$ close enough to $\boldsymbol{x}$. Second, the hyper-parameter λ could be quite sensitive to different test examples. Failing to choose a proper λ could also lead to failures



(a) Generated MNIST ($\epsilon = 1.0$) (b) Generated MNIST ($\epsilon = 2.0$) (c) Generated MNIST ($\epsilon = 3.0$)



(d) ImageNet10 ($\epsilon = 1.0$) (e) ImageNet10 ($\epsilon = 2.0$) (f) ImageNet10 ($\epsilon = 3.0$)

Figure 3.4: Comparisons between the theoretical intrinsic robustness bound and the empirically estimated unconstrained/in-distribution adversarial robustness, denoted as "unc" and "in" in the legend, of models produced during robust training on the generated data under ℓ_2. In each subfigure, the dotted curve line represents the theoretical bound on intrinsic robustness with horizontal axis denoting the different choice of α.

in finding in-distribution adversarial examples within ϵ constraint. In order to the tackle the aforementioned challenges, we propose to solve another optimization problem for the initialization of $\boldsymbol{z}$ and adopt binary search for the best choice of λ.

Figure 3.4 summarizes results from our empirical evaluations on intrinsic robustness of the generated MNIST and ImageNet10 data. We evaluate the empirical robustness of three types of robust training methods at different time points during the training procedure. To be more specific, we evaluate the robustness of the intermediate models produced every 5 training epochs. For each method, we plot both the unconstrained robustness measured by PGD attacks and the in-distribution robustness measured using the aforementioned strategies. In addition, based on the estimated local Lipschitz constants, we plot the implied theoretical bound on intrinsic robustness as the dotted line curve for direct comparison.

Compared with the intrinsic robustness upper bound (dotted curve line), the unconstrained robustness of various robustly-trained models is much smaller, and the gap between them becomes more obvious as we increase ϵ. However under all the considered settings, the estimated in-distribution adversarial robustness is much higher than the unconstrained one and closer to the theoretical upper bound, especially for the ImageNet10 data. Note that according to Remark 3.10, the actual intrinsic robustness $\overline{\text{AdvRob}}_\epsilon(\mathcal{F}_\alpha)$ should lie between the in-distribution robustness of any given classifier with risk at least α and the derived intrinsic robustness upper bound. Observing the big gap between the estimated in-distribution and unconstrained robustness of various robustly trained models, one would expect the current state-of-the-art robust models are still far from approaching the actual intrinsic robustness limit for real image distributions.

3.4 Concentration Estimation based Approach[5]

In Section 3.3, we present a method to understand intrinsic robustness using conditional generative models to connect the image space with the latent space whose concentration property is well-understood. However, one limitation is that the underlying input data is assumed to lie on the data manifold captured by some conditional generative model. That said, the results do not directly apply to the setting where the actual input distribution deviates from the assumed generated distribution. In this section, we are going to present an empirical method to directly measure the concentration function on a metric probability space, which can then be translated into an instrinsic robsutness limit using Theorem 3.5.

We aim to understand and empirically estimate the intrinsic robustness limit for typical

[5]Saeed Mahloujifar*, Xiao Zhang*, Mohammad Mahmoody, David Evans, *Empirically Measuring Concentration: Fundamental Limits on Intrinsic Robustness*, in the Thirty-third Conference on Neural Information Processing Systems (NeurIPS 2019) [81].

robust classification tasks by measuring concentration. Note that solving the concentration problem (3.1) itself only shows the existence of an error region $\mathcal{E}$ whose ϵ-expansion has certain (small) measure. This further implies the possibility of existing an optimally robust classifier (with risk at least α), whose robustness matches the intrinsic robustness limit $\overline{\text{AdvRob}}_\epsilon(\mathcal{F}_\alpha)$. However, actually finding such optimal classifier using a learning algorithm might be a much more challenging task.

3.4.1 Method for Measuring Concentration

There are two main challenges for solving the concentration of measure problem (3.1). First and foremost, we usually do not have access to the knowledge of the density function of the underlying distribution for typical robust classification tasks of interest. Moreover, even with the density function, solving the concentration problem (3.1) is still difficult, as we have to find the optimal subset among all the subsets within the whole search space.

We show how to overcome these challenges and find the actual concentration in the limit by first empirically simulating the distribution and then narrowing down our search space to a specific collection of subsets. Our results show that for a carefully chosen family of sets, the set with minimum expansion can be approximated using polynomially many samples. On the other hand, the minimum expansion convergence to the actual concentration (without the limits on the sets) as the complexity of the collection goes to infinity.

Before stating our main theorems, we introduce two useful definitions. The following definition captures the concentration function for a specific collection of subsets.

Definition 3.11 (Concentration Function for a Collection of Subsets)**.** Consider a metric probability space $(\mathcal{X}, \mu, \Delta)$. Let $\epsilon \geq 0$ and $\alpha \in (0, 1)$ be given parameters, then the con-

centration function with respect to a collection of subsets $\mathcal{G} \subseteq \mathsf{Pow}(\mathcal{X})$ is defined as

$$h(\mu, \alpha, \epsilon, \mathcal{G}) = \inf_{\mathcal{E} \in \mathcal{G}} \{\mu(\mathcal{E}_\epsilon) \colon \mu(\mathcal{E}) \geq \alpha\}.$$

Note that when $\mathcal{G} = \mathsf{Pow}(\mathcal{X})$, it corresponds to the standard concentration function $h(\mu, \alpha, \epsilon)$.

We also need to define the notion of complexity penalty for a collection of subsets. The complexity penalty for a collection of subsets captures the rate of the uniform convergence for the subsets in that collection. One can get such uniform convergence rates using the VC dimension or Rademacher complexity of the collection.

Definition 3.12 (Complexity Penalty)**.** Let $\mathcal{G} \subseteq \mathsf{Pow}(\mathcal{X})$ be a collection of subsets of $\mathcal{X}$. A function $\phi \colon \mathbb{N} \times \mathbb{R} \to [0, 1]$ is a complexity penalty for $\mathcal{G}$ iff for any probability measure μ supported on $\mathcal{X}$ and any $\delta \in [0, 1]$, we have

$$\Pr_{S \leftarrow \mu^m}[\exists\, \mathcal{E} \in \mathcal{G} \text{ s.t. } |\mu(\mathcal{E}) - \hat{\mu}_S(\mathcal{E})| \geq \delta] \leq \phi(m, \delta).$$

Theorem 3.13 shows how to overcome the challenge of measuring concentration from finite samples, when the concentration is defined with respect to specific families of subsets. Namely, it shows that the empirical concentration is close to the true concentration, if the underlying collection of subsets is not too complex.

Theorem 3.13 (Generalization of Concentration)**.** *Let $(\mathcal{X}, \mu, \Delta)$ be a metric probability space and $\mathcal{G} \subseteq \mathsf{Pow}(\mathcal{X})$. For any $\delta, \alpha, \epsilon \in [0, 1]$, we have*

$$\begin{aligned}\Pr_{S \leftarrow \mu^m}[h(\mu, \alpha - \delta, \epsilon, \mathcal{G}) - \delta \leq h(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G}) \leq & h(\mu, \alpha + \delta, \epsilon, \mathcal{G}) + \delta] \\ & \geq 1 - 2\big(\phi(m, \delta) + \phi_\epsilon(m, \delta)\big),\end{aligned}$$

where ϕ and ϕ_ϵ are complexity penalties for $\mathcal{G}$ and $\mathcal{G}_\epsilon$ respectively.

Proof of Theorem 3.13. Define $g(\mu, \alpha, \epsilon, \mathcal{G}) = \text{argmin}_{\mathcal{E}\in\mathcal{G}} \{\mu(\mathcal{E}_\epsilon) \colon \mu(\mathcal{E}) \geq \alpha\}$, and let $\mathcal{E} = g(\mu, \alpha+\delta, \epsilon, \mathcal{G})$ and $\hat{\mathcal{E}} = g(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G})$. (Note that these sets achieving the minimum might not exist, in which case we select a set for which the expansion is arbitrarily close to the infimum and every step of the proof will extend to this variant).

By the definition of the complexity penalty we have

$$\Pr_{S\leftarrow\mu^m}\left[\left|\mu(\hat{\mathcal{E}}) - \hat{\mu}_S(\hat{\mathcal{E}})\right| \geq \delta\right] \leq \phi(m, \delta),$$

which implies

$$\Pr_{S\leftarrow\mu^m}[\mu(\hat{\mathcal{E}}) \leq \alpha - \delta] \leq \phi(m, \delta).$$

Therefore, by the definition of h we have

$$\Pr_{S\leftarrow\mu^m}[\mu(\hat{\mathcal{E}}_\epsilon) \leq h(\mu, \alpha-\delta, \epsilon, \mathcal{G})] \leq \phi(m, \delta). \tag{3.12}$$

On the other hand, based on the definition of ϕ_ϵ we have

$$\Pr_{S\leftarrow\mu^m}\left[\left|\mu(\hat{\mathcal{E}}_\epsilon) - \hat{\mu}_S(\hat{\mathcal{E}}_\epsilon)\right| \geq \delta\right] \leq \phi_\epsilon(m, \delta). \tag{3.13}$$

Combining Equation 3.12 and Equation 3.13, and by a union bound we get

$$\Pr_{S\leftarrow\mu^m}[\hat{\mu}_S(\hat{\mathcal{E}}_\epsilon) \leq h(\mu, \alpha-\delta, \epsilon, \mathcal{G}) - \delta] \leq \phi(m, \delta) + \phi_\epsilon(m, \delta),$$

which by the definition of $\hat{\mathcal{E}}$ implies that

$$\Pr_{S\leftarrow\mu^m}[h(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G}) \leq h(\mu, \alpha-\delta, \epsilon, \mathcal{G}) - \delta] \leq \phi(m, \delta) + \phi_\epsilon(m, \delta). \tag{3.14}$$

Now we bound the probability for the other side of our inequality. By the definition of the notion of complexity penalty we have

$$\Pr_{S \leftarrow \mu^m} \left[\left| \mu(\mathcal{E}) - \hat{\mu}_S(\mathcal{E}) \right| \geq \delta \right] \leq \phi(m, \delta),$$

which implies

$$\Pr_{S \leftarrow \mu^m} [\hat{\mu}_S(\mathcal{E}) \leq \alpha] \leq \phi(m, \delta).$$

Therefore, by the definition of h we have,

$$\Pr_{S \leftarrow \mu^m} [\hat{\mu}_S(\mathcal{E}_\epsilon) \leq h(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G})] \leq \phi(m, \delta). \tag{3.15}$$

On the other hand, based on the definition of ϕ_ϵ we have

$$\Pr_{S \leftarrow \mu^m} \left[\left| \mu(\mathcal{E}_\epsilon) - \hat{\mu}_S(\mathcal{E}_\epsilon) \right| \geq \delta \right] \leq \phi(m, \delta) + \phi_\epsilon(m, \delta). \tag{3.16}$$

Combining Equations 3.15 and 3.16, by union bound we get

$$\Pr_{S \leftarrow \mu^m} [\mu(\mathcal{E}_\epsilon) \leq h(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G}) - \delta] \leq \phi(m, \delta) + \phi_\epsilon(m, \delta),$$

which by the definition of $\mathcal{E}$ implies

$$\Pr_{S \leftarrow \mu^m} [h(\mu, \alpha + \delta, \epsilon, \mathcal{G}) \leq h(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G}) - \delta] \leq \phi(m, \delta) + \phi_\epsilon(m, \delta). \tag{3.17}$$

Now combining Equations 3.14 and 3.17, by union bound we have

$$\begin{aligned}\Pr_{S\leftarrow\mu^m}[h(\mu,\alpha-\delta,\epsilon,\mathcal{G})-\delta\leq h(\hat{\mu}_S,\alpha,\epsilon,\mathcal{G})\leq h(\mu,\alpha+\delta,\epsilon,\mathcal{G})+\delta] \\ \geq 1-2\left(\phi(m,\delta)+\phi_\epsilon(m,\delta)\right),\end{aligned}$$

which completes the proof. □

Remark 3.14. Theorem 3.13 shows that if we narrow down our search to a collection of subsets $\mathcal{G}$ such that both $\mathcal{G}$ and $\mathcal{G}_\epsilon$ have small complexity penalty, then we can use the empirical distribution to measure concentration of measure for that specific collection. Note that the generalization bound of Theorem 3.13 depends on complexity penalties for both $\mathcal{G}$ and $\mathcal{G}_\epsilon$. Therefore, in order for this theorem to be useful, the collection $\mathcal{G}$ must be chosen in a careful way. For example, if $\mathcal{G}$ has bounded VC dimension, then $\mathcal{G}_\epsilon$ might still have a very large VC dimension. Alternatively, $\mathcal{G}$ might denote the collection of subsets that are decidable by a neural network of a certain size. In that case, even though there are well known complexity penalties for such collections (see [89]), the complexity of their expansions is unknown. In fact, relating the complexity penalty for expansion of a collection to that of the original collection is tightly related to generalization bounds in the adversarial settings, which has also been the subject of several recent works [22, 6, 85, 127, 98].

The following theorem states that if we gradually increase the complexity of the collection and the number of samples together, the empirical estimate of concentration converges to actual concentration, as long as several conditions hold. Theorem 3.15 and the techniques used in its proof are inspired by the work of [107] on learning minimum volume sets.

Theorem 3.15. *Let $\{\mathcal{G}(T)\}_{T\in\mathbb{N}}$ be a family of subset collections defined over a space $\mathcal{X}$. Let $\{\phi^T\}_{T\in\mathbb{N}}$ and $\{\phi^T_\epsilon\}_{T\in\mathbb{N}}$ be two families of complexity penalty functions such that ϕ^T and ϕ^T_ϵ are complexity penalties for $\mathcal{G}(T)$ and $\mathcal{G}_\epsilon(T)$ respectively, for some $\epsilon\in[0,1]$. Let*

$\{m(T)\}_{T\in\mathbb{N}}$ and $\{\delta(T)\}_{T\in\mathbb{N}}$ be two sequences such that $m(T) \in \mathbb{N}$ and $\delta(T) \in [0,1]$.

Consider a sequence of datasets $\{S_T\}_{T\in\mathbb{N}}$, where S_T consists of $m(T)$ i.i.d. samples from a measure μ supported on $\mathcal{X}$. Also let $\alpha \in [0,1]$ be such that h is locally continuous w.r.t the second parameter at point $(\mu, \alpha, \epsilon, \mathsf{Pow}(\mathcal{X}))$. If all the following hold,

1. $\sum_{T=1}^{\infty} \phi^T(m(T), \delta(T)) < \infty$
2. $\sum_{T=1}^{\infty} \phi_\epsilon^T(m(T), \delta(T)) < \infty$
3. $\lim_{T\to\infty} \delta(T) = 0$
4. $\lim_{T\to\infty} h(\mu, \alpha, \epsilon, \mathcal{G}(T)) = h(\mu, \alpha, \epsilon)$

then with probability 1, we have $\lim_{T\to\infty} h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{G}(T)) = h(\mu, \alpha, \epsilon)$.

Proof of Theorem 3.15. First, we lay out the following lemma which will be used in proving Theorem 3.15.

Lemma 3.16 (Borel-Cantelli Lemma)**.** *Let $\{E_T\}_{T\in\mathbb{N}}$ be a series of events such that*

$$\sum_{T=1}^{\infty} \Pr[E_T] < \infty$$

Then with probability 1, only finite number of events will occur.

Next, we prove Theorem 3.15. Define E_T to be the event that

$$h(\mu, \alpha - \delta(T), \epsilon, \mathcal{G}(T)) - \delta(T) > h(\hat{\mu}_{S_T}, \alpha, \epsilon) \text{ or}$$
$$h(\mu, \alpha + \delta(T), \epsilon, \mathcal{G}(T)) + \delta(T) < h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{G}).$$

Based on Theorem 3.13 we have $\Pr[E_T] \leq 2 \cdot (\phi^T(m(T), \delta(T)) + \phi_\epsilon^T(m(T), \delta(T)))$. Therefore, by Conditions 1 and 2 we have

$$\sum_{T=1}^{\infty} \Pr[E_T] \leq 2 \left(\sum_{T=1}^{\infty} \phi^T \left(m(T), \delta(T) \right) + \phi_\epsilon^T \left(m(T), \delta(T) \right) \right) < \infty.$$

Now by Lemma 3.16, we know there exist with measure 1 some $j \in \mathbb{N}$, such that for all $T \geq j$,

$$h(\mu, \alpha - \delta(T), \epsilon, \mathcal{G}(T)) - \delta(T) \leq h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{G}(T)) \leq h(\mu, \alpha + \delta(T), \epsilon, \mathcal{G}(T)) + \delta(T).$$

The above implies that

$$\begin{aligned} \lim_{T \to \infty} h(\mu, \alpha - \delta(T), \epsilon, \mathcal{G}(T)) - \delta(T) &\leq \lim_{T \to \infty} h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{G}(T)) \\ &\leq \lim_{T \to \infty} h(\mu, \alpha + \delta(T), \epsilon, \mathcal{G}(T)) + \delta(T). \end{aligned}$$

We know that

$$\begin{aligned} \lim_{T \to \infty} h(\mu, \alpha - \delta(T), \epsilon, \mathcal{G}(T)) &= \lim_{T_1 \to \infty} \lim_{T_2 \to \infty} h(\mu, \alpha - \delta(T_1), \epsilon, \mathcal{G}(T_2)) \\ \text{(By condition 4)} \quad &= \lim_{T_1 \to \infty} h(\mu, \alpha - \delta(T_1), \epsilon) \\ \text{(By local continuity and condition 3)} \quad &= h(\mu, \alpha, \epsilon). \end{aligned}$$

Similarly, we have

$$\lim_{T \to \infty} h(\mu, \alpha + \delta(T), \epsilon, \mathcal{G}(T)) = h(\mu, \alpha, \epsilon).$$

Therefore we have,

$$\lim_{T \to \infty} h(\mu, \alpha, \epsilon) - \delta(T) \leq \lim_{T \to \infty} h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{G}(T)) \leq \lim_{T \to \infty} h(\mu, \alpha, \epsilon) + \delta(T),$$

which by condition 3 implies

$$\lim_{T\to\infty} h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{G}(T)) = h(\mu, \alpha, \epsilon).$$

Thus, we complete the proof. □

Remark 3.17. In Theorem 3.15, the first two conditions restrict the growth rate for the complexity of the collections. Namely, we need the complexity penalties $\phi^T(m(T), \delta(T))$ and $\phi^T_\epsilon(m(T), \delta(T))$ to rapidly approach 0 as $T \to \infty$, which means the complexity of $\mathcal{G}(T)$ and $\mathcal{G}_\epsilon(T)$ should grow at a slow rate. The third condition requires that our generalization error goes to zero as we increase T. Note that the complexity penalty is a decreasing function with respect to δ, which means condition 3 makes achieving the first two conditions harder. However, since the complexity penalty is a function of both δ and sample size, we can still increase the sample size with a faster rate to satisfy the first two conditions. Finally, the fourth condition requires our approximation error goes to 0 as we increase T. Note that this condition holds for any family of collections of subsets that is a universal approximator (e.g., decision trees or neural networks). However, in order for our theorem to hold, we also need all the other conditions. In particular, we cannot use decision trees or neural networks as our collection of subsets, because we do not know if there is a complexity penalty for them that satisfies condition 2.

Special Case of ℓ_∞. In the following, we show how to instantiate Theorem 3.15 for the case of ℓ_∞-norm distance metric. Below, we introduce a special collection of subsets characterized by the *complement of a union of hyperrectangles*.

Definition 3.18 (Complement of union of hyperrectangles)**.** For any positive integer T, the collection of subsets specified by the *complement of a union of* T n*-dimensional hyperrect-*

angles is defined as:

$$\mathcal{CR}(T,n) = \left\{\mathbb{R}^n \setminus \cup_{t=1}^{T} \mathcal{R}ect(\boldsymbol{u}^{(t)}, \boldsymbol{r}^{(t)}) \colon \forall t \in [T], (\boldsymbol{u}^{(t)}, \boldsymbol{r}^{(t)}) \in \mathbb{R}^n \times \mathbb{R}^n_{\geq 0}\right\},$$

where $\mathcal{R}ect(\boldsymbol{u}, \boldsymbol{r}) = \left\{\boldsymbol{x} \in \mathcal{X} : \forall j \in [n], |x_j - u_j| \leq r_j/2\right\}$ denotes the hyperrectangle centered at $\boldsymbol{u}$ with $\boldsymbol{r}$ representing the edge size vector. When n is free of context, we simply write $\mathcal{CR}(T)$.

Recall that our goal is to find a subset $\mathcal{E} \in \mathbb{R}^n$ such that $\mathcal{E}$ has measure at least α and the ϵ_∞-expansion of $\mathcal{E}$ under ℓ_∞ has the minimum measure. To achieve this goal, we approximate the distribution μ with an empirical distribution $\hat{\mu}_S$, and limit our search to the special collection $\mathcal{CR}(T)$ (though our goal is to find the minimum concentration around arbitrary subsets). Namely, what we find is still an *upper bound* on the concentration function, and it is an upper bound that we know it converges the actual value in the limit. Our problem thus becomes the following optimization task:

$$\underset{\mathcal{E} \in \mathcal{CR}(T)}{\text{minimize}} \;\; \hat{\mu}_{\mathcal{S}}(\mathcal{E}_{\epsilon_\infty}) \quad \text{subject to } \; \hat{\mu}_{\mathcal{S}}(\mathcal{E}) \geq \alpha. \tag{3.18}$$

The following theorem provides the key to our empirical method by providing a convergence guarantee. It states that if we increase the number of rectangles and the number of samples together in a careful way, the solution to the problem using restricted sets converges to the true concentration.

Theorem 3.19. *Consider a nice metric probability space $(\mathbb{R}^n, \mu, \ell_\infty)$. Let $\{S_T\}_{T \in \mathbb{N}}$ be a family of datasets such that for all $T \in \mathbb{N}$, S_T contains at least T^4 i.i.d. samples from μ. For any ϵ_∞ and $\alpha \in [0,1]$, if h is locally continuous w.r.t the second parameter at point*

$(\mu, \alpha, \epsilon_\infty)$*, then with probability* 1 *we get*

$$\lim_{T\to\infty} h(\hat{\mu}_{S_T}, \alpha, \epsilon_\infty, \mathcal{CR}(T)) = h(\mu, \alpha, \epsilon_\infty).$$

Note that the size of S_T is selected as T^4 to guarantee conditions 1 and 2 are satisfied in Theorem 3.15. In fact, we can tune the parameters more carefully to get T^2, instead of T^4, but the convergence will be slower.

Proof of Theorem 3.19. This theorem follows from our general Theorem 3.15. We show that the choice of parameters here satisfies all four conditions of Theorem 3.15.

If we let $\mathcal{G}(T)$ to be the collection of subsets specified by complement of union of T hyperrectangles. Then $\mathcal{G}_\epsilon(T)$ will be the collection of of subsets specified by complement of union of T hyperrectangles that are bigger than ϵ in each coordinate. Therefore we have $\mathcal{G}_\epsilon(T) \subset \mathcal{G}(T)$. We know that the VC dimension of $\mathcal{G}(T)$ is $d_T = O(nT\log(T))$ because the VC dimension of all hyperrectangles is $O(n)$ and the functions formed by T fold union of functions in a VC class is at most $n \cdot T\log(T)$ (See [36]). Therefore, by VC inequality we have

$$\Pr_{S\leftarrow\mu^m}\left[\sup_{\mathcal{E}\in\mathcal{G}(T)} |\mu(\mathcal{E}) - \hat{\mu}_S(\mathcal{E})| \geq \delta\right] \leq 8e^{nT\log(T)\log(m) - m\delta^2/128}.$$

Therefore $\Phi^T(m,\delta) = 8e^{nT\log(T)\log(m)-m\delta^2/128}$ is a complexity penalty for both $\mathcal{G}(T)$ and $\mathcal{G}_\epsilon(T)$. Hence, if we define $\delta(T) = 1/T$ and $m(T) \geq T^4$, then the first three conditions of Theorem 3.15 are satisfied. The fourth condition is also satisfied by the universal consistency of histogram rules (See [28], Ch. 9). □

Special Case of ℓ_2. We demonstrate how to apply Theorem 3.15 to the case of ℓ_2. The following definition introduces the collection of subsets characterized by a *union of balls*:

Definition 3.20 (Union of Balls)**.** For any positive integer T, the collection of subsets specified by a *union of* T *n-dimensional balls* is defined as

$$\mathcal{B}(T, n) = \left\{ \cup_{t=1}^{T} \text{Ball}(\boldsymbol{u}^{(t)}, \boldsymbol{r}^{(t)}) \colon \forall t \in [T], (\boldsymbol{u}^{(t)}, \boldsymbol{r}^{(t)}) \in \mathbb{R}^n \times \mathbb{R}^n_{\geq 0} \right\}.$$

When n is free of context, we simply write $\mathcal{B}(T)$.

By restricting our search to the collection of a union of balls $\mathcal{B}(T)$ and replacing the underlying distribution μ with the empirical one $\hat{\mu}_{\mathcal{S}}$, our problem becomes the following optimization task

$$\underset{\mathcal{E} \in \mathcal{B}(T)}{\text{minimize}} \;\; \hat{\mu}_{\mathcal{S}}(\mathcal{E}_{\epsilon_2}) \quad \text{subject to } \hat{\mu}_{\mathcal{S}}(\mathcal{E}) \geq \alpha. \tag{3.19}$$

Theorem 3.21 guarantees that if we increase the number of balls and samples together in a careful way, the solution to the empirical problem (3.19) converges to the true concentration.

Theorem 3.21. *Consider a nice metric probability space* $(\mathbb{R}^n, \mu, \ell_2)$*. Let* $\{S_T\}_{T \in \mathbb{N}}$ *be a family of datasets such that for all* $T \in \mathbb{N}$*,* S_T *contains at least* T^4 *i.i.d. samples from* μ*. For any* ϵ_2 *and* $\alpha \in [0, 1]$*, if* h *is locally continuous w.r.t the second parameter at point* $(\mu, \alpha, \epsilon_2)$*, then with probability* 1 *we get*

$$\lim_{T \to \infty} h(\hat{\mu}_{S_T}, \alpha, \epsilon_2, \mathcal{B}(T)) = h(\mu, \alpha, \epsilon_2).$$

Proof of Theorem 3.21. Similar to Theorem 3.19 This theorem follows from our general Theorem 3.15. We show that the choice of parameters here satisfies all four conditions of Theorem 3.15.

If we let $\mathcal{G}(T)$ to be the collection of subsets specified by union of T balls. Then $\mathcal{G}_\epsilon(T)$ will

be the collection of of subsets specified by union of T balls with diameter at least ϵ. Similar to the proof of Theorem 3.19, we have $\mathcal{G}_\epsilon(T) \subset \mathcal{G}(T)$. We know that the VC dimension of all balls is $O(n)$ so using the fact that $\mathcal{G}(T)$ is T fold union of balls, the VC dimension of $\mathcal{G}(T)$ is $d_T = O(nT\log(T))$ (See [36]). Therefore, by VC inequality we have complexity penalties similar to those of Theorem 3.19 for both $\mathcal{G}(T)$ and $\mathcal{G}_\epsilon(T)$. Hence, if we define $\delta(T) = 1/T$ and $m(T) \geq T^4$, then the first three conditions of Theorem 3.15 are satisfied. The fourth condition is also satisfied by the universal consistency of kernel-based rules (See [28] , Ch. 10). □

3.4.2 Experiments for ℓ_∞

In this section, we provide heuristic methods to find the best possible error region, which covers at least α fraction of the samples and its expansion covers the least number of points for ℓ_∞ distance metric. Specifically, we first introduce our algorithm, then evaluate our approach on two benchmark image datasets: MNIST [73] and CIFAR-10 [68]. Note that in our experiments we exactly use the collection of subsets as suggested by our theoretical results in Section 3.4.1. However, that is not necessary and one might work with any subset collection to run experiments, as long as they can estimate the measure of the sets and their expansion. We tried working with other collection of subsets that we do not have theoretical support for (e.g. sets defined by a neural network) and observed a large generalization gap. This observation shows the importance of working with subset collections that we can theoretically control their generalization penalty.

Theorem 3.19 shows that the empirical concentration function $h(\hat{\mu}_{\mathcal{S}}, \alpha, \epsilon_\infty, \mathcal{CR}(T))$ converges to the actual concentration $h(\mu, \alpha, \epsilon_\infty)$ asymptotically, when T and $|\mathcal{S}|$ go to infinity

with $|\mathcal{S}| \geq T^4$. Thus, it remains to solve the empirical concentration problem (3.18).

Method. Although the collection of subsets is specified using simple topology, solving (3.18) exactly is still difficult, as the problem itself is combinatorial in nature. Borrowing techniques from clustering, we propose an empirical method to search for desirable error region within $\mathcal{CR}(T)$. Any error region $\mathcal{E}$ could be used to define $f_{\mathcal{E}}$, i.e., $f_{\mathcal{E}}(\boldsymbol{x}) = c(\boldsymbol{x}),$ if $\boldsymbol{x} \notin \mathcal{E};\ f_{\mathcal{E}}(\boldsymbol{x}) \neq c(\boldsymbol{x}),$ if $\boldsymbol{x} \in \mathcal{E}$. However, finding a classifier corresponding to $f_{\mathcal{E}}$ using a learning algorithm might be a difficult task. Here, we find the optimally robust error region, not the corresponding classifier. A desirable error region should have small adversarial risk[6], compared with all subsets in $\mathcal{CR}(T)$ that have measure at least α.

The high-level intuition is that images from different classes are likely to be concentrated in separable regions, since it is generally believed that small perturbations preserve the ground-truth class at the sampled images. Therefore, if we cluster all the images into different clusters, a desired region with low adversarial risk should exclude any image from the dense clusters, otherwise the expansion of such a region will quickly cover the whole cluster. In other words, a desirable subset within $\mathcal{CR}(T)$ should be ϵ_∞ away (in ℓ_∞ norm) from all the dense image clusters, which motivates our method to cover the dense image clusters using hyperrectangles and treat the complement of them as error set.

More specifically, our algorithm (for pseudocode, see Algorithm 2) starts by sorting all the training images in an ascending order based on the ℓ_1-norm distance to the k-th nearest neighbour with $k = 50$, and then obtains T hyperrectangular image clusters by performing k-means clustering [54] on the top-q densest images, where the metric is chosen as ℓ_1 and the maximum iterations is set as 30. Finally, we perform a binary search over $q \in [0, 1]$, where we set $\delta_{\text{bin}} = 0.005$ as the stopping criteria, to obtain the best robust subset (lowest

[6]The adversarial risk of an error region $\mathcal{E}$ simply refers to the adversarial risk of $f_{\mathcal{E}}$.

Algorithm 2: Heuristic Search for Robust Error Region under ℓ_∞

Input : a set of images $\mathcal{S}$; perturbation strength ϵ_∞; error threshold α; number of hyperrectangles T; number of nearest neighbours k; precision for binary search δ_{bin}.

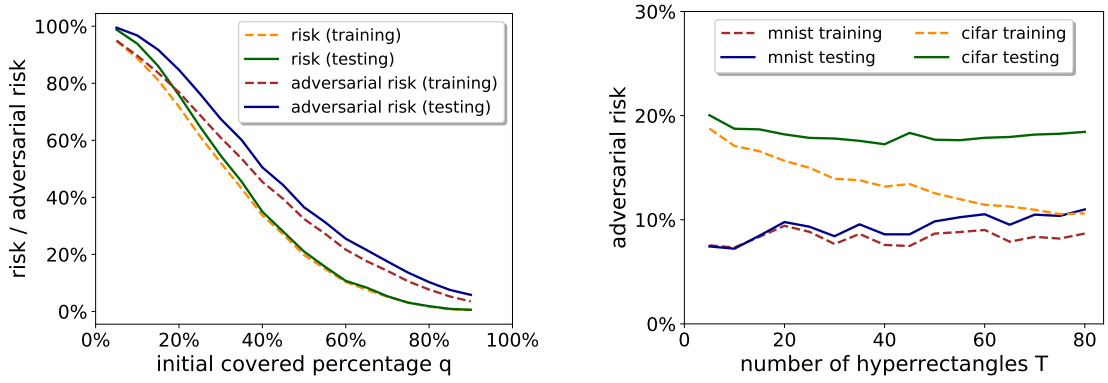
$r_k(\boldsymbol{x}) \leftarrow$ compute the ℓ_1-norm distance to the k-th nearest neighbour for each $\boldsymbol{x} \in \mathcal{S}$;
$\mathcal{S}_{\text{sort}} \leftarrow$ sort all the images in $\mathcal{S}$ by $r_k(\boldsymbol{x})$ in an ascending order;
$q_{\text{lower}} \leftarrow 0.0, \quad q_{\text{upper}} \leftarrow 1.0$;
while $q_{upper} - q_{lower} > \delta_{bin}$ **do**
    $q \leftarrow (q_{\text{lower}} + q_{\text{upper}})/2$;
    perform kmeans clustering algorithm (T clusters, ℓ_1 metric) on the top-q images of $\mathcal{S}_{\text{sort}}$;
    $\{\boldsymbol{u}^{(t)}\}_{t=1}^{T} \leftarrow$ record the centroids of the resulted T clusters;
    for $t = 1, 2, \ldots, T$ **do**
        $\mathcal{R}ect(\boldsymbol{u}^{(t)}, \boldsymbol{r}^{(t)}) \leftarrow$ cover t-th cluster with the minimum-sized rectangle centered at $\boldsymbol{u}^{(t)}$;
    end
    $\mathcal{E}_q \leftarrow \mathcal{X} \setminus \cup_{t=1}^{T} \mathcal{R}ect_{\epsilon_\infty}(\boldsymbol{u}^{(t)}, \boldsymbol{r}^{(t)})$; // $\mathcal{R}ect_\epsilon(\boldsymbol{u}, \boldsymbol{r})$ denotes the ϵ-expansion of $\mathcal{R}ect(\boldsymbol{u}, \boldsymbol{r})$
    if $|\mathcal{S} \cap \mathcal{E}_q|/|\mathcal{S}| \geq \alpha$ **then**
        $q_{\text{lower}} \leftarrow q, \quad AdvRisk_q \leftarrow \left|\left\{\boldsymbol{x} \in \mathcal{S} : \boldsymbol{x} \notin \cup_{t=1}^{T} \mathcal{R}ect(\boldsymbol{u}^{(t)}, \boldsymbol{r}^{(t)})\right\}\right|/|\mathcal{S}|$;
    else
        $q_{\text{upper}} \leftarrow q$;
    end
end
$\hat{q} \leftarrow \text{argmin}_q\{AdvRisk_q\}$;

Output : $(\hat{q}, AdvRisk_{\hat{q}}, \mathcal{E}_{\hat{q}})$

adversarial risk) in $\mathcal{CR}(T)$ with empirical measure at least α.

Results. We choose α to reflect the best accuracy achieved by state-of-the-art classifiers, using $\alpha = 0.01$ and $\epsilon_\infty \in \{0.1, 0.2, 0.3, 0.4\}$ for MNIST and selecting appropriate values to represent the best typical results on the other datasets (see Table 3.4). Given the number of hyperrectangles, T, we obtain the resulting error region using the proposed algorithm on the training dataset, and tune T for the minimum adversarial risk on the testing dataset.

Figure 3.5 shows the learning curves regarding risk and adversarial risk for two specific experimental settings. Figure 3.5(a) suggests that as we increase the initial covered percentage



(a) varying q (b) varying T

Figure 3.5: (a) Plots of risk and adversarial risk w.r.t. the resulted error region using our method as q varies (CIFAR-10, $\epsilon_\infty = 8/255$, $T = 30$); (b) Plots of adversarial risk w.r.t. the resulted error region using our method (best q) as T varies on MNIST ($\epsilon_\infty = 0.3$) and CIFAR-10 ($\epsilon_\infty = 8/255$).

q, both risk and adversarial risk of the corresponding error region decrease. This supports our use of binary search on q in Algorithm 2. On the other hand, as can be seen from Figure 3.5(b), overfitting with respect to adversarial risk becomes significant as we increase the number of hyperrectangles. According to the adversarial risk curve for testing data, the optimal value of T is selected as $T = 10$ for MNIST ($\epsilon_\infty = 0.3$) and $T = 40$ for CIFAR-10 ($\epsilon_\infty = 8/255$).

Table 3.4 summarizes the optimal parameters, the empirical risk and adversarial risk of the learned error region on the testing datasets for each experimental setting. Since the k-means algorithm does not guarantee global optimum, we repeat our method for 10 runs with random restarts in terms of the best parameters, then report both the mean and the standard deviation. Our experiments provide examples of rather robust error regions for real image datasets. For instance, in Table 3.4 we have a case where the measure of the resulting error region increases from 5.94% to 18.13% after expansion with $\epsilon_\infty = 8/255$ on CIFAR-10 dataset. This means that there could potentially be a classifier with 5.94% risk and 18.13% adversarial risk, but the-state-of-the-art robust classifier has empirically-

Table 3.4: Summary of the main results using our method with ℓ_∞ perturbations.

Dataset	α	ϵ_∞	T	**Best** q	**Risk** (%)	**AdvRisk** (%)
MNIST	0.01	0.1	5	0.662	1.23 ± 0.12	3.64 ± 0.30
		0.2	10	0.660	1.11 ± 0.10	5.89 ± 0.44
		0.3	10	0.629	1.15 ± 0.13	7.24 ± 0.38
		0.4	10	0.598	1.21 ± 0.09	9.92 ± 0.60
CIFAR-10	0.05	2/255	10	0.680	5.72 ± 0.25	8.13 ± 0.26
		4/255	20	0.688	6.05 ± 0.40	13.66 ± 0.33
		8/255	40	0.734	5.94 ± 0.34	18.13 ± 0.30
		16/255	75	0.719	5.28 ± 0.23	28.83 ± 0.46

measured adversarial risk 52.96% [79].

Noticing that the risk lower threshold $\alpha = 0.05$ is much lower than the empirical risk 12.70% of the adversarially-trained robust model reported in [79], we further measure the empirical concentration on MNIST and CIFAR-10 using our method with α set to be the same as the reported standard test error in [79], which is demonstrated in Table 3.5. In particular, we show that the gap between the attack success rate of Madry et al.'s classifier (10.70%) and our estimated best-achievable adversarial risk (8.28%) is quite small on MNIST, suggesting that the robustness of Madry et al.'s classifier is actually close to the intrinsic robustness. In sharp contrast, the gap becomes significantly larger on CIFAR-10: 29.21% for our estimate, while 52.96% for the reported attack success rate in [79]. Regardless of the difference, this gap cannot be explained by the concentration of measure phenomenon, suggesting there may still be room for developing more robust classifiers, or that other inherent reasons impede learning a more robust classifier.

Table 3.5: Comparisons between our method and the existing adversarially trained robust classifiers under different settings. We use the *Risk* and *AdvRisk* for robust training methods to denote the standard test error and attack success rate reported in literature. The *AdvRisk* reported for our method can be seen as an estimated lower bound of adversarial risk for existing classifiers.

Dataset	**Strength (metric)**	**Method**	**Risk**	**AdvRisk**
MNIST	$\epsilon_\infty = 0.3$	[79]'s	1.20%	10.70%
		Ours	$1.35\% \pm 0.08\%$	$8.28\% \pm 0.22\%$
MNIST	$\epsilon_2 = 1.5$	[106]'s	1.00%	20.00%
		Ours	1.08%	2.12%
CIFAR-10	$\epsilon_\infty = 8/255$	[79]'s	12.70%	52.96%
		Ours	$14.22\% \pm 0.46\%$	$29.21\% \pm 0.35\%$

3.4.3 Experiments for ℓ_2

For ℓ_2 adversaries, Theorem 3.21 guarantees the asymptotic convergence of the empirical concentration function characterized by union of balls $\mathcal{B}(T)$ towards the actual concentration. Thus, it remains to solve the corresponding optimization problem (3.19). Similar to ℓ_∞, we propose an empirical method to search for desirable robust error regions under ℓ_2 perturbations. From a high level, our algorithm (for pseudocode, see Algorithm 3) places T balls in a sequential manner, and searches for the best possible placement using a greedy approach at each time. Since enumerating all the possible ball centers is infeasible, we restrict the choice of the center to be the set of training data points. Our method keeps two sets of indices: one for the initial coverage and one for the coverage after expansion, and updates them when we find the optimal placement, i.e. the ball centered at some training data point that has the minimum expansion with respect to both sets.

We compare our empirical method for finding robust error regions characterized by a union of balls with the hyperplane-based approach [49] on MNIST and CIFAR-10. In particular,

Algorithm 3: Heuristic Search for Robust Error Region under ℓ_2

Input : a set of images $\mathcal{S}$; perturbation strength ϵ_2; error threshold α; number of balls T.

$\hat{\mathcal{E}} \leftarrow \{\}, \quad \hat{\mathcal{S}}_{\text{init}} \leftarrow \{\}, \quad \hat{\mathcal{S}}_{\text{exp}} \leftarrow \{\}$;

for $t = 1, 2, \ldots, T$ **do**

  $k_{\text{lower}} \leftarrow \lceil (\alpha|\mathcal{S}| - |\hat{\mathcal{S}}_{\text{init}}|)/(T - t + 1) \rceil, \quad k_{\text{upper}} \leftarrow (\alpha|\mathcal{S}| - |\hat{\mathcal{S}}_{\text{init}}|)$;

  for $\boldsymbol{u} \in \mathcal{S}$ **do**

    for $k \in [k_{lower}, k_{upper}]$ **do**

      $r_k(\boldsymbol{u}) \leftarrow$ compute the ℓ_2 distance from $\boldsymbol{u}$ to the k-th nearest neighbour in $\mathcal{S} \setminus \hat{\mathcal{S}}_{\text{init}}$;

      $\mathcal{S}_{\text{init}}(\boldsymbol{u}, k) \leftarrow \{\boldsymbol{x} \in \mathcal{S} \setminus \hat{\mathcal{S}}_{\text{init}} : \|\boldsymbol{x} - \boldsymbol{u}\|_2 \leq r_k(\boldsymbol{u})\}$;

      $\mathcal{S}_{\text{exp}}(\boldsymbol{u}, k) \leftarrow \{\boldsymbol{x} \in \mathcal{S} \setminus \hat{\mathcal{S}}_{\text{exp}} : \|\boldsymbol{x} - \boldsymbol{u}\|_2 \leq r_k(\boldsymbol{u}) + \epsilon_2\}$;

    end

  end

  $(\hat{\boldsymbol{u}}, \hat{k}) \leftarrow \text{argmin}_{(\boldsymbol{u},k)}\{|\mathcal{S}_{\text{exp}}(\boldsymbol{u}, k)| - |\mathcal{S}_{\text{init}}(\boldsymbol{u}, k)|\}$;

  $\hat{\mathcal{E}} \leftarrow \hat{\mathcal{E}} \cup \text{Ball}(\hat{\boldsymbol{u}}, r_{\hat{k}}(\hat{\boldsymbol{u}}))$;

  $\hat{\mathcal{S}}_{\text{init}} \leftarrow \hat{\mathcal{S}}_{\text{init}} \cup \mathcal{S}_{\text{init}}(\hat{\boldsymbol{u}}, \hat{k}), \quad \hat{\mathcal{S}}_{\text{exp}} \leftarrow \hat{\mathcal{S}}_{\text{exp}} \cup \mathcal{S}_{\text{exp}}(\hat{\boldsymbol{u}}, \hat{k})$;

end

Output : $\hat{\mathcal{E}}$

the risk threshold α is set to be the same as the case of ℓ_∞, and the adversarial strength ϵ_2 is chosen such that the volume of an ℓ_2 ball with radius ϵ_2 is roughly the same as the ℓ_∞ ball with radius ϵ_∞, using the conversion rule $\epsilon_2 = \sqrt{n/\pi} \cdot \epsilon_\infty$ as in [122]. Table 3.6 summarizes the optimal parameters, the testing risk and adversarial risk of the trained error regions using different methods, where we tune the number of balls T for our method.

Our results show that there exist rather robust ℓ_2 error regions for real image datasets. For example, the measure of the resulting error region using our method only increases by 0.69% (from 5.14% to 5.83%) after expansion with $\epsilon_2 = 0.4905$ on CIFAR-10. Compared with [49], our method is able to find regions with significantly smaller adversarial risk (around half the adversarial risk of regions found by their method) on MNIST, while attaining comparable error region robustness on CIFAR-10. Nevertheless, the adversarial risk attained by state-of-the-art robust classifiers against ℓ_2 perturbations is much higher than these reported

Table 3.6: Comparisons between different methods for finding robust regions with ℓ_2 metric.

Dataset	α	ϵ_2	**[49]**'s Method		**Our Method**		
			Risk	AdvRisk	T	Risk	AdvRisk
MNIST	0.01	1.58	1.18%	3.92%	20	1.07%	2.19%
		3.16	1.18%	9.73%	20	1.02%	4.15%
		4.74	1.18%	23.40%	20	1.07%	10.09%
CIFAR-10	0.05	0.2453	5.27%	5.58%	5	5.16%	5.53%
		0.4905	5.27%	5.93%	5	5.14%	5.83%
		0.9810	5.27%	6.47%	5	5.12%	6.56%

rates (see Table 3.5 for a comparison with the best robust classifier against ℓ_2 perturbations proposed in [106]).

3.5 Improved Concentration Estimation using Half Spaces[7]

In Section 3.4, we present an empirical method to measure the concentration of an arbitrary distribution using data samples, then employed it to estimate a lower bound on intrinsic robustness for image benchmarks. By demonstrating the gap between the estimated bounds of intrinsic robustness and the robustness performance achieved by the best current models, we show that concentration of measure is not the sole reason behind the adversarial vulnerability of existing classifiers for benchmark image distributions. However, due to the heuristic nature of the proposed algorithm, it remains elusive whether the estimates it produces can serve as useful approximations of the underlying intrinsic robustness limits, thus hindering understanding of how much of the actual adversarial risk can be explained by the

[7]Jack Prescott, Xiao Zhang, David Evans, *Improved Estimation of Concentration Under Lp-Norm Distance Metric Using Half Spaces*, in the Ninth International Conference on Learning Representations (ICLR 2021) [96].

concentration of measure phenomenon.

In this section, we address this issue by first characterizing the optimum of the actual concentration problem for general Gaussian spaces, then using our theoretical insights to develop an alternative algorithm for measuring concentration empirically that significantly improves both the accuracy and efficiency of estimates of intrinsic robustness. While we do not demonstrate a specific classifier which achieves this robustness upper bound, our results rule out inherent image distribution concentration as the reason for our current inability to find adversarially robust models.

3.5.1 Generalizing the Gaussian Isoperimetric Inequality

Before proceeding to introduce the proposed methodology for solving the concentration of measure problem, we first present our main theoretical results of generalizing the Gaussian Isoperimetric Inequality. This theoretical result largely motivates our method.

Note that the Gaussian Ispoperimetric Inequality (see Lemma 3.4) characterizes the optimum of the concentration problem (3.1) with respect to standard Gaussian distribution and ℓ_2-distance, where half spaces are proven to be the optimal sets.

Definition 3.22 (Half Space)**.** Let $\boldsymbol{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Without loss of generality, assume $\|\boldsymbol{w}\|_2 = 1$. An n-dimensional *half space* with parameters $\boldsymbol{w}$ and b is defined as:

$$\mathcal{H}_{\boldsymbol{w},b} = \{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{w}^\top \boldsymbol{z} + b \leq 0\}.$$

Lemma 3.4 implies the concentration function with respect spherical Gaussian distribution and ℓ_2-norm distance metric. However, it only gives a concentration function for estimating the the intrinsic robustness limit in a very restrictive setting. To understand the concentra-

tion of measure for more general problems, we prove the following theorem that extends the standard Gaussian Isoperimetric Inequality (Lemma 3.4) to non-spherical Gaussian measure and general ℓ_p-norm distance metrics for any $p \geq 2$.

Theorem 3.23 (Generalized Gaussian Isoperimetric Inequality)**.** *Let ν be the probability measure of $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\theta} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}$ is a positive definite matrix in $\mathbb{R}^{n\times n}$. Consider the probability space $(\mathbb{R}^n, \nu)$ with ℓ_p-norm distance, where $p \geq 2$ (including ℓ_∞). For any $\mathcal{E} \in \text{Pow}(\mathbb{R}^n)$ and $\epsilon \geq 0$,*

$$\nu\big(\mathcal{E}_\epsilon^{(\ell_p)}\big) \geq \Phi\big(\Phi^{-1}\big(\nu(\mathcal{E})\big) + \epsilon/\|\boldsymbol{\Sigma}^{1/2}\|_p\big), \tag{3.20}$$

where $\boldsymbol{\Sigma}^{1/2}$ is the square root of $\boldsymbol{\Sigma}$, and $\|\boldsymbol{\Sigma}^{1/2}\|_p$ denotes the induced matrix p-norm of $\boldsymbol{\Sigma}^{1/2}$.

Proof of Theorem 3.23. We provide the proof sketch of Theorem 3.23 as follows. The complete proof of the theorem can be found in [96]. We start with the spherical Gaussian distribution where $\nu = \gamma_n$. More specifically, we are going to prove that for any $\mathcal{E} \subseteq \mathbb{R}^n$ and $\eta \geq 0$,

$$\gamma_n\big(\mathcal{E}_\eta^{(\ell_p)}\big) \geq \Phi\big(\Phi^{-1}\big(\gamma_n(\mathcal{E})\big) + \eta\big) \text{ holds for } p \geq 2. \tag{3.21}$$

Note that for any vector $\boldsymbol{x} \in \mathbb{R}^n$, the mapping $p \to \|\boldsymbol{x}\|_p$ is monotonically decreasing for any $p \geq 1$ (see [100]), thus we can show that $\mathcal{E}_\eta^{(\ell_q)} \subseteq \mathcal{E}_\eta^{(\ell_p)}$ holds for any $p \geq q \geq 1$. Making use of the standard Gaussian Isoperimetric Inequality (Lemma 3.4), we then immediately obtain

$$\gamma_n\big(\mathcal{E}_\eta^{(\ell_p)}\big) \geq \gamma_n\big(\mathcal{E}_\eta^{(\ell_2)}\big) \geq \Phi\big(\Phi^{-1}\big(\gamma_n(\mathcal{E})\big) + \eta\big), \text{ for any } p \geq 2.$$

Moreover, to prove the concentration bound for general case where ν is the probability

measure of $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, we build connections with the spherical Gaussian case by constructing a subset $\mathcal{A} = \{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\theta}) : \boldsymbol{x} \in \mathcal{E}\}$. Based on the affine transformation of Gaussian measure, we then prove:

$$\nu(\mathcal{E}) = \gamma_n(\mathcal{A}) \quad \text{and} \quad \nu(\mathcal{E}_\epsilon^{(\ell_p)}) \geq \gamma_n(\mathcal{A}_\eta^{(\ell_p)}), \quad \text{where } \eta = \epsilon / \|\boldsymbol{\Sigma}^{1/2}\|_p. \tag{3.22}$$

Finally, combining (3.21) and (3.22) completes the proof of Theorem 3.23. □

Remark 3.24. Theorem 3.23 suggests that for general Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and any ℓ_p-norm distance ($p \geq 2$), the corresponding concentration function is lower bounded by $\Phi(\Phi^{-1}(\alpha) + \epsilon / \|\boldsymbol{\Sigma}^{1/2}\|_p)$. Due to the NP-hardness of approximating the matrix p-norm [57], it is generally hard to infer whether the equality of (3.20) can be attained or not. However, for specific special Gaussian spaces, we can derive optimal subsets that achieve the lower bound. In particular, for the case where $\boldsymbol{\Sigma} = \mathbf{I}_n$ and $p > 2$, the optimum is attained when $\mathcal{E}$ is a half space with axis-aligned weight vector (that is, $\boldsymbol{w} = \boldsymbol{e}_j$ for some $j \in [n]$). For the case where $\boldsymbol{\Sigma} \neq \mathbf{I}_n$ and $p = 2$, the optimal solution is a half space $\mathcal{H}_{\boldsymbol{v}_1, b}$, where $\boldsymbol{v}_1$ is the eigenvector with respect to the largest eigenvalue of $\boldsymbol{\Sigma}$.

Proof of the Optimality Results in Remark 3.24. First, we prove the optimality for the spherical Gaussian case, where $\nu = \gamma_n$ and $p > 2$. Let $\mathcal{H} = \mathcal{H}_{\boldsymbol{w}, b}$ be a half space with axis-aligned weight vector, that said $\boldsymbol{w} = \boldsymbol{e}_j$ for some $j \in [n]$. Intuitively speaking, the ϵ-expansion of $\mathcal{H}$ with respect to ℓ_p-norm will only happen along the j-th dimension. More rigorously, we are going to prove the following results: for any $\epsilon \geq 0$,

$$\mathcal{H}_\epsilon^{(\ell_p)} = \mathcal{H}_\epsilon^{(\ell_2)} \text{ holds for any } p \geq 1. \tag{3.23}$$

By definition, $\mathcal{H} = \{\boldsymbol{x} \in \mathbb{R}^n : x_j + b \leq 0\}$. For any $\boldsymbol{x} \notin \mathcal{H}$, let $\hat{\boldsymbol{x}} \in \mathcal{H}$ be the closest point of $\boldsymbol{x}$ in terms of ℓ_p-norm. Since the weight vector $\boldsymbol{w}$ of $\mathcal{H}$ is axis-aligned, thus $\hat{\boldsymbol{x}}$ will only

differ from $\boldsymbol{x}$ by the j-th element. That said, $\hat{x}_{j'} = x_{j'}$ for any $j' \neq j$ and $\hat{x}_j = -b$. Thus for any $p \geq 1$, we have $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_p = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 = x_j + b$. Based on this observation, we further obtain that for any $p \geq 1$,

$$\mathcal{H}_\epsilon^{(\ell_p)} = \{\boldsymbol{x} \in \mathbb{R}^n : x_j + b \leq \epsilon\} = \mathcal{H}_\epsilon^{(\ell_2)},$$

which proves (3.23). According to the Gaussian Isoperimetric Inequality (Lemma 3.4), we obtain

$$\gamma_n\big(\mathcal{H}_\epsilon^{(\ell_p)}\big) = \gamma_n\big(\mathcal{H}_\epsilon^{(\ell_2)}\big) = \Phi(\Phi^{-1}(\gamma_n(\mathcal{H})) + \epsilon).$$

Therefore, combining this with Theorem 3.23, we prove the optimality for the spherical Gaussian case.

Now we turn to prove the non-spherical Gaussian case with $p = 2$. Based on Theorem 3.23, the lower bound is $\Phi(\Phi^{-1}(\nu(\mathcal{E}) + \epsilon/\|\boldsymbol{\Sigma}^{1/2}\|_2)$ when $p = 2$. In the following, we are going to prove: if we choose $\mathcal{E} = \mathcal{H}_{\boldsymbol{v}_1,b}$, where $\boldsymbol{v}_1$ is the eigenvector with respect to the largest eigenvalue of $\boldsymbol{\Sigma}$, this lower bound is attained. Similarly to the proof of Theorem 3.23, we construct $\mathcal{A} = \{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\theta}) : \boldsymbol{x} \in \mathcal{E}\}$.

Note that when $\mathcal{E}$ is a half space, the constructed set $\mathcal{A}$ is also a half space. In particular, for the case where $\mathcal{E} = \mathcal{H}_{\boldsymbol{v}_1,b}$, for any $\boldsymbol{u} \in \mathcal{A}$, there exists an $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{u} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\theta})$ and $\boldsymbol{v}_1^\top \boldsymbol{x} + b \leq 0$. This implies that $\boldsymbol{v}_1^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{u} + \boldsymbol{v}_1^\top \boldsymbol{\theta} + b \leq 0$ for any $\boldsymbol{u} \in \mathcal{A}$. Since $\boldsymbol{v}_1$ is the eigenvector of $\boldsymbol{\Sigma}$, we further have that $\mathcal{A}$ is a half space with weight vector $\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_1 = \|\boldsymbol{\Sigma}^{1/2}\|_2 \cdot \boldsymbol{v}_1$.

Note that according to (3.22), as in the proof of Theorem 3.23, for any $\mathcal{E} \subseteq \mathbb{R}^n$, we have

$$\nu(\mathcal{E}) = \gamma_n(\mathcal{A}) \text{ and } \nu\big(\mathcal{E}_\epsilon^{(\ell_2)}\big) \geq \gamma_n\big(\mathcal{A}_\eta^{(\ell_2)}\big), \text{ where } \eta = \epsilon/\|\boldsymbol{\Sigma}^{1/2}\|_2.$$

For $\mathcal{E} = \mathcal{H}_{\boldsymbol{v}_1,b}$, based on the explicit formulation of ℓ_2-distance to a half space, we can explicitly compute the η-expansion of $\mathcal{A}$ as

$$\mathcal{A}_\eta^{(\ell_2)} = \{\boldsymbol{u} \in \mathbb{R}^n : \boldsymbol{v}_1^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{u} + \boldsymbol{v}_1^\top \boldsymbol{\theta} + b \le \eta \cdot \|\boldsymbol{\Sigma}^{1/2}\|_2\}.$$

When we set $\eta = \epsilon/\|\boldsymbol{\Sigma}^{1/2}\|_2$, it further implies that

$$\gamma_n\big(\mathcal{A}_\eta^{(\ell_2)}\big) = \Pr_{\boldsymbol{u}\sim\gamma_n}\big[\boldsymbol{v}_1^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{u} + \boldsymbol{v}_1^\top \boldsymbol{\theta} + b \le \epsilon\big] = \Pr_{\boldsymbol{x}\sim\nu}\big[\boldsymbol{v}_1^\top \boldsymbol{x} + b \le \epsilon\big] = \nu\big(\mathcal{E}_\epsilon^{(\ell_2)}\big).$$

Finally, according to the optimality of the standard Gaussian Isoperimetric Inequality (Lemma 3.4), we complete the proof. □

3.5.2 Empirically Measuring Concentration using Half Spaces

Built upon our concentration estimation method developed in Section 3.4, we consider the following empirical counterpart of the actual concentration problem (3.1):

$$\underset{\mathcal{E}\in\mathcal{G}}{\text{minimize}}\ \hat{\mu}_m\big(\mathcal{E}_\epsilon^{(\ell_p)}\big) \quad \text{subject to}\ \ \hat{\mu}_m(\mathcal{E}) \ge \alpha, \tag{3.24}$$

where $\hat{\mu}_m$ is the empirical measure based on $\{\boldsymbol{x}_i\}_{i\in[m]}$ and $\mathcal{G} \subseteq \text{Pow}(\mathcal{X})$ denotes a particular collection of subsets. Previously, we use the complement of union of T hyperrectangles as $\mathcal{G}$ for ℓ_∞ and the union of T balls for ℓ_2, and prove that if one increases the complexity parameter T and the sample size m together in a careful way, the optimal value of the empirical concentration problem (3.24) converges to the actual concentration asymptotically. However, it is unclear how quickly it converges and how well the proposed heuristic algorithm finds the optimum of (3.24).

We argue that the set of half spaces is a superior choice for $\mathcal{G}$ with respect to any ℓ_p-norm distance. Apart from achieving the optimality for certain Gaussian spaces as discussed in Remark 3.24, estimating concentration using half spaces has several other advantages including the closed-form solution of ℓ_p-distance to half-space (Lemma 3.25) and its small sample complexity requirement for generalization (Theorem 3.26). To be more specific, we focus on the following optimization problem based on the empirical measure $\hat{\mu}_m$ and the collection of half spaces $\mathcal{HS}(n)$:

$$\underset{\mathcal{E}\in\mathcal{HS}(n)}{\text{minimize}}\ \hat{\mu}_m\big(\mathcal{E}_\epsilon^{(\ell_p)}\big) \quad \text{subject to}\ \ \hat{\mu}_m(\mathcal{E}) \geq \alpha, \tag{3.25}$$

where $\mathcal{HS}(n) = \{\mathcal{H}_{\boldsymbol{w},b} : \boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}, \text{ and } \|\boldsymbol{w}\|_2 = 1\}$ is the set of half spaces in $\mathbb{R}^n$. In the following discussions of this section, we write $\mathcal{E}_\epsilon^{(\ell_p)} = \mathcal{E}_\epsilon$ for simplicity.

The following lemma characterizes the closed-form solution of the ℓ_p-norm distance between a point $\boldsymbol{x}$ and a half space. Such a formulation enables an exact computation of the empirical measure with respect to the ϵ-expansion of any half space.

Lemma 3.25 (ℓ_p-Distance to Half Space)**.** *Let $\mathcal{H}_{\boldsymbol{w},b} \in \mathcal{HS}(n)$ be an n-dimensional half space. For any vector $\boldsymbol{x} \in \mathbb{R}^n$, the ℓ_p-norm distance ($p \geq 1$) from $\boldsymbol{x}$ to $\mathcal{H}_{\boldsymbol{w},b}$ is:*

$$d_p(\boldsymbol{x}, \mathcal{H}_{\boldsymbol{w},b}) = \begin{cases} 0, & \boldsymbol{w}^\top \boldsymbol{x} + b \leq 0; \\ (\boldsymbol{w}^\top \boldsymbol{x} + b)/\|\boldsymbol{w}\|_q, & \textit{otherwise}. \end{cases}$$

Here, q is a real number that satisfies $1/p + 1/q = 1$.

Proof of lemma 3.25. We only consider the case when $\boldsymbol{w}^\top \boldsymbol{x} + b > 0$, because $d_p(\boldsymbol{x}, \mathcal{H}_{\boldsymbol{w},b})$ is zero trivially holds if $\boldsymbol{w}^\top \boldsymbol{x} + b \leq 0$. The problem of finding the ℓ_p-distance from a given point $\boldsymbol{x}$ to a half space $\mathcal{H}_{\boldsymbol{w},b}$ can be formulated as the following constrained optimization

problem:

$$\min_{\boldsymbol{z}\in\mathbb{R}^n} \|\boldsymbol{z}-\boldsymbol{x}\|_p, \quad \text{subject to } \boldsymbol{w}^\top \boldsymbol{z}+b\leq 0. \tag{3.26}$$

Let $\tilde{\boldsymbol{z}}=\boldsymbol{z}-\boldsymbol{x}$, then optimization problem (3.26) is equivalent to

$$\min_{\tilde{\boldsymbol{z}}\in\mathbb{R}^n} \|\tilde{\boldsymbol{z}}\|_p, \quad \text{subject to } \boldsymbol{w}^\top \tilde{\boldsymbol{z}}+\boldsymbol{w}^\top \boldsymbol{x}+b\leq 0. \tag{3.27}$$

According to Hölder's Inequality, for any $\tilde{\boldsymbol{z}}\in\mathbb{R}^n$ we have

$$-\|\boldsymbol{w}\|_q\cdot\|\tilde{\boldsymbol{z}}\|_p\leq \boldsymbol{w}^\top\tilde{\boldsymbol{z}}\leq \|\boldsymbol{w}\|_q\cdot\|\tilde{\boldsymbol{z}}\|_p,$$

where $1/p+1/q=1$. Therefore, for any $\tilde{\boldsymbol{z}}$ that satisfies the constraint of (3.27), we have

$$\boldsymbol{w}^\top\boldsymbol{x}+b\leq -\boldsymbol{w}^\top\tilde{\boldsymbol{z}}\leq \|\boldsymbol{w}\|_q\cdot\|\tilde{\boldsymbol{z}}\|_p. \tag{3.28}$$

Since $\|\boldsymbol{w}\|_2=1$, we have $\|\boldsymbol{w}\|_q>0$, thus (3.28) further suggests $\|\tilde{\boldsymbol{z}}\|_p\geq(\boldsymbol{w}^\top\boldsymbol{x}+b)/\|\boldsymbol{w}\|_q$.

Up till now, we have proven that the optimal value of (3.26) is lower bounded by $(\boldsymbol{w}^\top\boldsymbol{x}+b)/\|\boldsymbol{w}\|_q$. The remaining task is to show this lower bound can be achieved. To this end, we construct $\hat{\boldsymbol{z}}$ as

$$\hat{z}_j=x_j-\frac{\boldsymbol{w}^\top\boldsymbol{x}+b}{\|\boldsymbol{w}\|_q}\cdot\left(\frac{\boldsymbol{w}_j^q}{\sum_{j\in[n]}\boldsymbol{w}_j^q}\right)^{1/p}, \quad \text{for any } j\in[n],$$

where $1/p+1/q=1$. We remark that for the extreme case where $p=\infty$, such choice of $\hat{\boldsymbol{z}}$ can be simplified as $\hat{\boldsymbol{z}}=\boldsymbol{x}-(\boldsymbol{w}^\top\boldsymbol{x}+b)\cdot\text{sgn}(\boldsymbol{w})/\|\boldsymbol{w}\|_q$, where $\text{sgn}(\cdot)$ denotes the sign

function for vectors. According to the construction, it can be verified that

$$\boldsymbol{w}^\top \hat{\boldsymbol{z}} + b = (\boldsymbol{w}^\top \boldsymbol{x} + b) - \frac{\boldsymbol{w}^\top \boldsymbol{x} + b}{\|\boldsymbol{w}\|_q} \cdot \sum_{j\in[n]} w_j \cdot \left(\frac{\boldsymbol{w}_j^q}{\sum_{j\in[d]} \boldsymbol{w}_j^q}\right)^{1/p} = 0,$$

and $\|\hat{\boldsymbol{z}} - \boldsymbol{x}\|_p = (\boldsymbol{w}^\top \boldsymbol{x} + b)/\|\boldsymbol{w}\|_q$. □

Lemma 3.25 implies that the ϵ-expansion of any half space with respect to the ℓ_p-norm is still a half space. Since the VC-dimensions of both the set of half spaces and its expansion are bounded, we can thus apply Theorem 3.13, which yields the following theorem that characterizes the generalization of concentration with respect to half spaces.

Theorem 3.26 (Generalization of Concentration of Half Spaces)**.** *Consider the metric probability space, $(\mathcal{X}, \mu, \|\cdot\|_p)$, where $\mathcal{X} \subseteq \mathbb{R}^n$ and $p \geq 1$. Let $\{\boldsymbol{x}_i\}_{i\in[m]}$ be a set of m instances sampled from μ, and let $\hat{\mu}_m$ be the corresponding empirical measure. Define the concentration functions regarding the collection of half spaces $\mathcal{HS}(n)$ with respect to μ as:*

$$h\big(\mu, \alpha, \epsilon, \mathcal{HS}(n)\big) = \inf_{\mathcal{E}\in\mathcal{HS}(n)} \{\mu(\mathcal{E}_\epsilon) \colon \mu(\mathcal{E}) \geq \alpha\},$$

and let $h(\hat{\mu}_m, \alpha, \epsilon, \mathcal{HS}(n))$ be its empirical counterpart with respect to $\hat{\mu}_m$. For any $\delta \in (0, 1)$, there exists constants c_0 and c_1 such that with probability at least $1 - c_0 \cdot e^{-n\log n}$,

$$h(\hat{\mu}_m, \alpha - \delta, \epsilon, \mathcal{HS}(n)) - \delta \leq h(\hat{\mu}_m, \alpha, \epsilon, \mathcal{HS}(n)) \leq h\big(\hat{\mu}_m, \alpha + \delta, \epsilon, \mathcal{HS}(n)\big) + \delta$$

holds, provided that the sample size $m \geq c_1 \cdot n\log n/\delta^2$.

Proof of Theorem 3.26. We write $\mathcal{HS}$ as $\mathcal{HS}(n)$ for simplicity. Let S be a set of size m sampled from μ and $\hat{\mu}_m$ be the corresponding empirical measure. Note that the VC-

dimension of $\mathcal{HS}(n)$ is $n+1$ (see [84]), thus according to the VC inequality, we have

$$\Pr_{S\leftarrow\mu^m}\Big[\sup_{\mathcal{E}\in\mathcal{HS}(n)}\big|\hat{\mu}_m(\mathcal{E})-\mu(\mathcal{E})\big|\geq\delta\Big]\leq 8e^{(n+1)\log(m+1)-m\delta^2/32}.$$

In addition, according to Lemma 3.25, the ϵ-expansion of any half space is still a half space. Therefore, we can directly apply Theorem 3.3 in [81] to bound the generalization of concentration with respect to half spaces: for any $\delta\in(0,1)$, we have

$$\Pr_{S\leftarrow\mu^m}\Big[h(\hat{\mu}_m,\alpha-\delta,\epsilon,\mathcal{HS})-\delta\leq h(\hat{\mu}_m,\alpha,\epsilon,\mathcal{HS})\leq h(\hat{\mu}_m,\alpha+\delta,\epsilon,\mathcal{HS})+\delta\Big]$$
$$\geq 1-32e^{(n+1)\log(m+1)-m\delta^2/32}.$$

Finally, assuming the sample size $m\geq c_0\cdot n\log n/\delta^2$ for some constant c_0 large enough, then there exists positive constant c_1 such that

$$h(\hat{\mu}_m,\alpha-\delta,\epsilon,\mathcal{HS})-\delta\leq h(\hat{\mu}_m,\alpha,\epsilon,\mathcal{HS})\leq h(\hat{\mu}_m,\alpha+\delta,\epsilon,\mathcal{HS})+\delta$$

holds with probability at least $1-c_1\cdot e^{-n\log n}$. □

Remark 3.27. Theorem 3.26 suggests that for the concentration of measure problem with respect to half spaces, in order to achieve δ estimation error with high probability, it requires $\Omega(n\log(n)/\delta^2)$ number of samples. Compared with [81], our method using half spaces requires fewer samples in theory to achieve the same estimation error.[8] For standard Gaussian inputs, the empirical concentration with respect to half spaces is guaranteed to converge to the actual concentration as in (3.1), i.e., $\lim_{m\to\infty}h(\hat{\mu}_m,\alpha,\epsilon,\mathcal{HS}(n))=h(\hat{\mu}_m,\alpha,\epsilon)$; whereas for distributions that are not Gaussian, there might exist a gap. However, this gap of empirical and actual concentration is shown to be uniformly small across various data

[8]The proposed estimators for ℓ_∞ and ℓ_2 in [81] require $\Omega(nT\log(n)\log(T)/\delta^2)$ samples to achieve δ approximation, where T is a predefined number of hyperrectangles or balls.

distributions, as will be discussed in our experiments

Based on Lemma 3.25, estimating the empirical concentration using half spaces as defined in (3.25) is equivalent to solving the following constrained optimization problem:

$$\begin{aligned} \underset{\boldsymbol{w}\in\mathbb{R}^n, b\in\mathbb{R}}{\text{minimize}} \quad & \sum_{i\in[m]} \mathbb{1}\{\boldsymbol{w}^\top \boldsymbol{x}_i + b \le \epsilon \|\boldsymbol{w}\|_q\} \\ \text{subject to} \quad & \frac{1}{m}\sum_{i\in[m]} \mathbb{1}\{\boldsymbol{w}^\top \boldsymbol{x}_i + b \le 0\} \ge \alpha \ \text{ and } \ \|\boldsymbol{w}\|_2 = 1. \end{aligned} \tag{3.29}$$

The optimal solution to (3.29) would be a half space $\mathcal{H}_{\boldsymbol{w},b}$ that satisfies the following two properties: (1) approximately α-fraction of data is covered by $\mathcal{H}_{\boldsymbol{w},b}$, and (2) most of the remaining data points are at least ϵ-away from $\mathcal{H}_{\boldsymbol{w},b}$ under ℓ_p-norm distance metric.

Note that we can always set b to be the α-quantile of the projections $\{-\boldsymbol{w}^\top \boldsymbol{x}_i : i \in [m]\}$ to satisfy the first property. In addition, to satisfy the second condition, inspired by the special case optimality results in Remark 3.24, we propose to search for a weight vector $\boldsymbol{w}$ such that both the ℓ_q-norm of $\boldsymbol{w}$ is small and the variation of the given sample set along the direction of $\boldsymbol{w}$ is large. These searching criteria guarantee that the given dataset $\{\boldsymbol{x}_i\}_{i\in[m]}$, when projected onto $\boldsymbol{w}$ then normalized by $\|\boldsymbol{w}\|_q$, will have a large variance, which implies the second property.

We propose a heuristic algorithm to search for the desirable half space according to the aforementioned criteria. In particular, Algorithm 4 searches for a desirable half space based on the principal components of the empirical dataset and their rotations defined by a power parameter. More specifically, the function pow() takes a vector $\boldsymbol{v} \in \mathbb{R}^n$ and a positive integer $s \in \mathbb{Z}^+$, and returns the normalized s-th power of $\boldsymbol{v}$ (with sign preserved):

$$\text{pow}(\boldsymbol{v}, s) = \text{sgn}(\boldsymbol{v}) \circ [\text{abs}(\boldsymbol{v})]^s / \|\boldsymbol{v}^s\|_2 = \begin{cases} \boldsymbol{v}^s/\|\boldsymbol{v}^s\|_2, & \text{if } s \text{ is odd;} \\ \text{sgn}(\boldsymbol{v}) \circ \boldsymbol{v}^s/\|\boldsymbol{v}^s\|_2, & \text{otherwise.} \end{cases} \tag{3.30}$$

Algorithm 4: Heuristic Search for Robust Half Space under ℓ_p-distance

Input : a set of samples $\{\boldsymbol{x}_i\}_{i\in[m]}$; strength ϵ (in ℓ_p-norm); risk threshold α; #iterations S.

$\mathbf{Q} \leftarrow$ compute the sample covariance matrix based on $\{\boldsymbol{x}_i\}_{i\in[m]}$;

$\mathcal{V} \leftarrow$ obtain the set of principal components by eigenvalue decomposition on $\mathbf{Q}$;

for $\boldsymbol{v} \in \mathcal{V}$ **do**

  for $s = 1, 2, \ldots, S$ **do**

    $\boldsymbol{w} \leftarrow$ select from $\{\pm\text{pow}(\boldsymbol{v}, s)\}$; // pow() is defined according to (3.30)

    $b \leftarrow \alpha$-quantile of the set $\{-\boldsymbol{w}^\top \boldsymbol{x}_i : i \in [m]\}$;

    $\text{AdvRisk}_\epsilon(\mathcal{H}_{\boldsymbol{w},b}) \leftarrow \sum_{i=1}^m \mathbb{1}(\boldsymbol{w}^\top \boldsymbol{x}_i + b \leq \epsilon\|\boldsymbol{w}\|_q)/m$;

  end

end

$(\hat{\boldsymbol{w}}, \hat{b}) \leftarrow \text{argmin}_{(\boldsymbol{w},b)} \text{AdvRisk}_\epsilon(\mathcal{H}_{\boldsymbol{w},b})$;

Output : $\mathcal{H}_{\hat{\boldsymbol{w}},\hat{b}}$

Note that all the functions used in (3.30) are element-wise operations for vectors, where $\text{sgn}(\boldsymbol{v})$, $\text{abs}(\boldsymbol{v})$, $\boldsymbol{v}^s$ represent the sign, absolute value and the s-th power of $\boldsymbol{v}$ respectively, and the operator $\circ$ denotes the Hardamard product of two vectors.

Connected with the theoretical optimum regarding Gaussian spaces in Remark 3.24, the top principal component corresponds to the optimal choice of $\boldsymbol{w}$ if the perturbation metric is ℓ_2-distance, whereas close-to-axis would be favourable for $\boldsymbol{w}$ when $p > 2$. In addition, as implied by the empirical concentration problem (3.29) and the monotonicity of ℓ_p-mapping, the value of $\|\boldsymbol{w}\|_q$ will be more influential in affecting the ϵ-expansion of half space as p grows larger. For example, the ℓ_∞-norm of $\boldsymbol{w}$ can be as large as $\sqrt{n}$ for the worst case (n denotes the input dimension), while $\|\boldsymbol{w}\|_\infty = 1$ if $\boldsymbol{w}$ aligns any axis. By searching through the region between each principal component and the closest axis, the proposed algorithm aims to find the optimal balance between $\|\boldsymbol{w}\|_q$ and the variance of the given data along $\boldsymbol{w}$ that leads to the smallest ϵ-expansion. Although there is no theoretical guarantee that our algorithm will find the optimum to (3.29) for an arbitrary dataset, we empirically show its

efficacy in our experiments in estimating concentration across various datasets.

Moreover, our algorithm is efficient in terms of both time and space complexities. Precomputing the principal components requires $O(mn^2+n^3)$ time and $O(n^2)$ space to store them, where m denotes the samples size and n is the input dimension. For each iteration step, the time complexity of computing $\boldsymbol{w}, b$ and $\text{AdvRisk}_\epsilon(\mathcal{H}_{\boldsymbol{w},b})$ is $O(mn)$, while the space complexity for saving the intermediate variables and the best parameters is $O(m+n)$. With n outer iterations and S inner iterations, the total time complexity is $O(n^3+mn^2S)$. The total space complexity is $O(n^2+mn)$, where the extra $O(mn)$ denotes the initial space requirement for saving all the input data. For our experiments, we observe $\text{AdvRisk}_\epsilon(\mathcal{H}_{\boldsymbol{w},b})$ is not sensitive to small increment of the exponent parameter s, thus we choose to increase s in a more aggressive way, which further saves computation.

3.5.3 Experiments

In this section, we evaluate our empirical method for estimating concentration under ℓ_∞-norm distance and comparing its performance to that of the method proposed in Section 3.4. We first demonstrate that the estimate produced by our algorithm is very close to the actual concentration for a spherical Gaussian distribution, and that our method is able to find much tighter bounds on the best possible adversarial risk for several image benchmarks. We then compare the convergence rates, and show that our method converges with substantially less data. Note that while we only provide results for the most widely-used ℓ_∞-norm perturbation metric adopted in the existing adversarial examples literature, our algorithm and experiments can be applied to any other ℓ_p-norm.

Estimation Accuracy. First, we evaluate the performance of our algorithm under ℓ_∞-norm distance metric on a generated synthetic dataset consisting of 30,000 samples from

$\mathcal{N}(\mathbf{0}, \mathbf{I}_{784})$. Since the proposed method follows from the analytical results of concentration of multivariate Gaussian distributions, we expect results produced by our empirical method to closely approach the analytical concentration on this simulated Gaussian dataset. We initially consider the case where $\epsilon = 1.0$ and $\alpha = 0.5$ for the actual concentration problem, requiring that the feasible set contains at least half of the data samples, and the adversary can perturb each entry by precisely the standard deviation of the underlying distribution. Our algorithm is able to produce a half space whose ϵ-expansion has mean empirical measure 84.18% over 5 repeated trials. According to Theorem 3.23 and Remark 3.24, the optimal value of the considered concentration problem is 84.13%. This implies that our method performs very well when the underlying distribution is Gaussian, while in stark contrast, the method proposed in Section 3.4 is not able to find a region whose expansion has measure less than 1 on the same simulated set. In addition, we consider another setting for this dataset where $\epsilon = 1.0$ is set the same and $\alpha = 0.05$ is set to be much smaller. Similarly, we observe that our method significantly outperforms the previous method in terms of the estimation accuracy (see Table 3.7 for the detailed comparison results).

Next, we evaluate our method on several image benchmarks. We set the values of α and ϵ to be the same as in Section 3.4 for the ℓ_∞ case. For example, we use $\alpha = 0.01,\ \epsilon \in \{0.1, 0.2, 0.3, 0.4\}$ for MNIST, and $\alpha = 0.05,\ \epsilon \in \{2/255, 4/255, 8/255, 16/255\}$ for CIFAR-10. These α values were selected to roughly represent the standard error of the state-of-the-art classifiers.

Table 3.7 demonstrates the risk and adversarial risk with respect to the best produced subsets using both methods, computed on a separate test dataset. In our context of measuring concentration, risk refers to the empirical measure of the produced subset, while adversarial risk corresponds to the empirical measure of its ϵ-expansion. We use a $50/50$ train-test split over the whole dataset to perform our evaluation, and determine the best exponent of each

Table 3.7: Comparisons between our method of estimating concentration with ℓ_∞-norm distance and the method proposed by [81] for different settings. For $\mathcal{N}(\mathbf{0}, \mathbf{I}_{784})$ with $\alpha = 0.5$ and $\epsilon = 1.0$, the previous method is unable to produce nontrivial estimate. Results for the previous method are taken directly from the original paper (except for the Gaussian results).

Dataset	$\boldsymbol{\alpha}$	$\boldsymbol{\epsilon}$	**Test Risk (%)**		**Test Adv. Risk (%)**	
			Prev. Method	Our Method	Prev. Method	Our Method
$\mathcal{N}(\mathbf{0}, \mathbf{I}_{784})$	0.05	1.0	6.21 ± 0.44	5.20 ± 0.16	89.75 ± 0.80	26.37 ± 0.17
	0.5	1.0	-	49.98 ± 0.25	-	84.18 ± 0.11
MNIST	0.01	0.1	1.23 ± 0.12	1.22 ± 0.05	3.64 ± 0.30	1.35 ± 0.06
		0.2	1.11 ± 0.10	1.24 ± 0.05	5.89 ± 0.44	1.52 ± 0.06
		0.3	1.15 ± 0.13	1.25 ± 0.04	7.24 ± 0.38	1.75 ± 0.05
		0.4	1.21 ± 0.09	1.27 ± 0.05	9.92 ± 0.60	1.98 ± 0.08
CIFAR-10	0.05	$2/255$	5.72 ± 0.25	5.14 ± 0.13	8.13 ± 0.26	5.28 ± 0.12
		$4/255$	6.05 ± 0.40	5.22 ± 0.20	13.66 ± 0.33	5.68 ± 0.21
		$8/255$	5.94 ± 0.34	5.22 ± 0.16	18.13 ± 0.30	6.28 ± 0.13
		$16/255$	5.28 ± 0.23	5.19 ± 0.08	28.83 ± 0.46	7.34 ± 0.15
FMNIST	0.05	0.1	5.92 ± 0.85	5.33 ± 0.14	11.56 ± 0.84	6.04 ± 0.13
		0.2	6.00 ± 1.02	5.34 ± 0.14	14.82 ± 0.71	6.82 ± 0.19
		0.3	6.13 ± 0.93	5.24 ± 0.10	17.46 ± 0.53	8.01 ± 0.19
SVHN	0.05	0.01	8.83 ± 0.30	5.23 ± 0.09	10.17 ± 0.29	5.56 ± 0.08

principal component based on a brute-force search. Though our method is deterministic for a given pair of training and testing sets, we account for the variance of our method over different train-test splits by repeating our experiments 5 times and reporting the mean and standard deviation of the results for each (α, ϵ). It is worth noting that the randomness of the previous method is derived not only from the selection of the training and test sets, but also from the inherent randomness of the employed k-means algorithm.

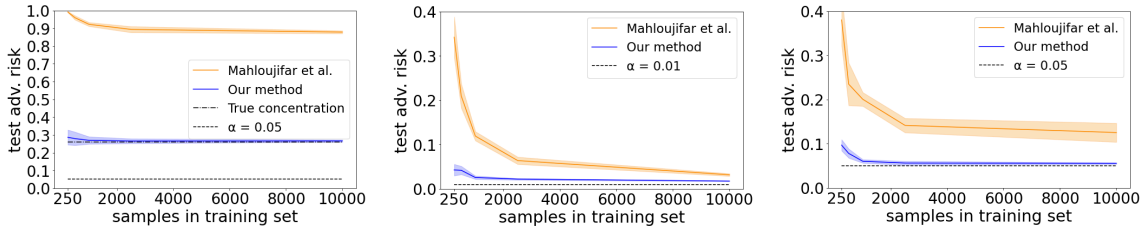
We observe from Table 3.7 that in every case, the estimated adversarial risk is significantly lower for our method than for the one found by the previous method. Since both methods restrict the search space to some special collection of subsets, these estimates can be viewed as valid empirical upper bounds of the actual concentration as defined in (3.1). Therefore,

the fact that our results are significantly lower indicates that our algorithm is able to produce estimates that are much closer to the optimum of the targeted problem. In addition, when translated to adversarial robustness, these tighter estimates prove the existence of a rather robust classifier[9] that has risk at least α, which further suggests that the underlying intrinsic robustness limit of each of these image benchmarks is actually much higher than previously thought.

For example, the best classifier produced by the previous method using hyperrectangles has 18.1% adversarial risk under ℓ_∞-perturbations bounded by $\epsilon = 8/255$ on CIFAR-10. However, our results demonstrate that the adversarial risk of the best possible robust classifier can be as low as 6.3% given the same risk constraint, indicating the underlying intrinsic robustness to be above 93.7%. As the intrinsic robustness limits are shown to be very close to the trivial upper bound $1 - \alpha$ across all the settings, our results reveal that the concentration of measure phenomenon is not an important factor that causes the adversarial vulnerability of existing classifiers on these image benchmarks.

Convergence Analysis. Figure 3.6 shows the convergence rate of our method under ℓ_∞-distance for Gaussian data from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{784})$ $(\alpha = 0.05,\ \epsilon = 1)$, as well as for MNIST $(\alpha = 0.01,\ \epsilon = 0.1)$ and CIFAR-10 $(\alpha = 0.05,\ \epsilon = 2/255)$. For each graph, the horizontal x-axis represents the size of the dataset used to train the estimator, and the vertical y-axis shows the concentration bounds estimated for a separate test set, which is of size $30,000$ for each case. We generate the means and standard deviations for these convergence curves by repeating both methods 5 times for different randomly-selected training and test tests. For the method presented in Section 3.4, we tune the number of the hyperrectangles T for the optimal performance based on the empirically-observed adversarial risk.

[9]Based on the ground-truth c and the returned set $\mathcal{E}$ of our algorithm, this classifier can be simply constructed by setting $f(\boldsymbol{x}) = c(\boldsymbol{x})$ for $\boldsymbol{x} \notin \mathcal{E}$ and $f(\boldsymbol{x}) \neq c(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{E}$. Without knowing the ground-truth, we note that such classifier may or may not be learnable.



(a) $\mathcal{N}(\mathbf{0}, \mathbf{I}_{784})$ (b) MNIST (c) CIFAR-10

Figure 3.6: The convergence curves of the best possible adversarial risk estimated using our method and the previous method as the number of training samples grows.

For the simulated Gaussian datasets, we include a horizontal line at $y = 0.2595$ to represent the true concentration of the underlying distribution, derived from Theorem 3.23 and Remark 3.24. This allows us to more accurately assess the convergence of our method, as it is the only case for which we know the optimal value that our empirical estimates should be converging to. We see that our estimates approach this line very quickly, coming within 0.01 of the true value given only about 1,000 samples.

While we do not have such a theoretical limit for other datasets, the risk threshold α can be viewed as a lower bound of the actual concentration, which is useful in visually assessing the convergence performance of our method. We can see that for both MNIST and CIFAR-10, our estimates get very close to the horizontal line at $y = \alpha$ given several thousand training samples. Since the actual concentration must be no less than α and our estimated upper bound is approaching α from the above, we immediate infer that the actual concentration of these data distributions with ℓ_∞-norm distance should be a value sightly greater than α. These results not only demonstrate the superiority of our method over the previous method in estimating concentration, but also show that concentration of measure is not the reason for our inability to find adversarially robust models for these image benchmarks.

3.6 Discussion

In the previous sections, we proposed different methods for characterizing the intrinsic robustness limits (with respect to imperfect classifiers) for typical classification tasks on image benchmarks. This section provides further discussions on these methods as well as their connections with existing literature.

Comparisons between our Methods. Following the line of related works [44, 69], our method presented in Section 3.3 first trains a conditional generative model using samples from the underlying data distribution, then estimates an upper bound on the intrinsic robustness limit with respect to the generated distribution based on the smoothness parameter of generator. One limitation of this method is the estimated concentration property is with respect to the simulated distribution captured by the learned generator instead of the actual data distribution. It is possible to generalize our intrinsic robustness bound by introducing an extra term regarding the Wasserstein distance between the generated distribution and the actual distribution (see Theorem 4 in [44] for a similar adaptation), however, there will be an empirical challenge on how to get good estimates of such Wasserstein distance. In addition, since the empirical estimates are with respect to an upper bound on intrinsic robustness, it remains unknown how to rigorously characterize the gap between the actual intrinsic robustness limit and the estimated upper bounds. We made an initial attempt to estimate the in-distribution adversarial risk of state-of-the-art adversarially trained models to empirically understand such gap (see Section 3.3.3 for the experimental results on this).

Different from generative model based approaches, our methods proposed in Sections 3.4 and 3.5 directly learns the concentration function of the underlying distribution. Then, we employ the learned concentration function to estimate intrinsic robustness limits with respect to imperfect classifiers. These methods are related to the previous work [49], which

provides experiments to heuristically estimate the concentration property of MNIST dataset under ℓ_2 distance. In comparison, our method in Section 3.4 provides a general methodology to empirically estimate the concentration of measure with provable guarantees, and is able to deal with ℓ_∞ perturbations, which is the most popular setting for research in adversarial examples; while our method in Section 3.5 generalizes the approach in [49] to any ℓ_p-norm distance metric with $p \geq 2$ and is able to produce tighter empirical concentration estimates for image benchmarks.

From the theoretical perspective, both our concentration estimation based methods proposed in Sections 3.4 and 3.5 adopt the same empirical framework based on carefully-chosen collection of subsets but with different choices (union of hyperrectangles or balls are used in Section 3.4 and half spaces in Section 3.5). When the concentration problem is restricted to those selected collections of subsets, generalization of concentration results can be proved (see Theorem 3.13 and Theorem 3.26). However, we additionally proved a convergence guarantee of the concentration with respect to the collection of subsets used in Section 3.4 to the actual concentration (see Theorem 3.15). We are not able to obtain such convergence guarantee using half spaces, since they are not universal set approximators.

On the algorithmic side, both the proposed algorithms in Sections 3.4 and 3.5 for solving the empirical concentration problem are able to return an optimally-found 'robust' subset within the selected collection of subsets, which is then used as the error region of the best robust classifier we are able to construct. The adversarial robustness of such constructed classifier is regarded as an empirical estimate of the intrinsic robustness limit. We demonstrate in Section 3.5 that choosing the set of half spaces for empirical concentration estimation is superior than previous choices in producing tighter estimates for measuring concentration

of several benchmark datasets.

Error Analysis for Concentration Estimation. At the core of our concentration estimation approaches is selecting a good collection of subsets $\mathcal{G}$ for the empirical concentration estimation problem. In the following, we first conduct a detailed error analysis for the concentration estimators proposed in Sections 3.4 and 3.5, then discuss its implications for measuring concentration of general metric probability spaces.

Let $(\mathcal{X}, \mu, \Delta)$ be the underlying input metric probability space and $\mathcal{G} \in \text{Pow}(\mathcal{X})$ be a collection of subsets. Consider the following empirical concentration problem:

$$\underset{\mathcal{E} \in \mathcal{G}}{\text{minimize}} \ \hat{\mu}_m(\mathcal{E}_\epsilon) \quad \text{subject to} \ \hat{\mu}_m(\mathcal{E}) \geq \alpha, \tag{3.31}$$

where $\hat{\mu}_m$ is the empirical measure based on m i.i.d. samples from μ, and $\alpha \in (0, 1)$, $\epsilon \geq 0$ are given constants. Suppose an algorithm aims to solve problem (3.31) and returns $\mathcal{E}$ as a solution. The approximation error between the empirical estimate of concentration and the actual concentration can be decomposed into three error terms:

$$\begin{aligned} \underbrace{h(\mu, \alpha, \epsilon) - \hat{\mu}_m(\mathcal{E}_\epsilon)}_{\text{approximation error}} = \underbrace{h(\mu, \alpha, \epsilon) - h(\mu, \alpha, \epsilon, \mathcal{G})}_{\text{modeling error}} + \underbrace{h(\mu, \alpha, \epsilon, \mathcal{G}) - h(\hat{\mu}_m, \alpha, \epsilon, \mathcal{G})}_{\text{finite sample estimation error}} \\ + \underbrace{h(\hat{\mu}_m, \alpha, \epsilon, \mathcal{G}) - \hat{\mu}_m(\mathcal{E}_\epsilon)}_{\text{optimization error}}. \end{aligned} \tag{3.32}$$

The *modeling error* denotes the difference between the actual concentration function and the concentration function with respect to the selected collection of subsets $\mathcal{G}$; the *finite sample estimation error* represents the generalization gap between the empirical concentration function and its limit; and the *optimization error* captures how well the algorithm approximates the empirical concentration problem. Such an error decomposition applies to

both the empirical methods proposed in both Sections 3.4 and 3.5.

The complexity of $\mathcal{G}$ and the complexity of its ϵ-expansion $\mathcal{G}_\epsilon = \{\mathcal{E}_\epsilon : \mathcal{E} \in \mathcal{G}\}$ control the finite sample estimation error. So, $\mathcal{G}$ should be selected such that the empirical concentration function $h(\hat{\mu}_m, \alpha, \epsilon, \mathcal{G})$ generalizes. If either $\mathcal{G}$ or $\mathcal{G}_\epsilon$ is too complex (e.g., it has unbounded VC-dimension), it will be difficult to control the generalization of the empirical concentration function.

There exist tradeoffs among the three error terms in (3.32), and it is unlikely that there is a uniformly good choice for $\mathcal{G}$ that minimizes all these error terms. In particular, increasing the complexity of $\mathcal{G}$ typically reduces the modeling error, since the feasible set of the concentration function with respect to $\mathcal{G}$ becomes larger. However, according to the generalization of concentration, this will also increase the finite sample estimation error. Therefore, we should consider the effect of all these error terms when choosing $\mathcal{G}$, including the hardness of the optimization problem with respect to the empirical concentration. It is favorable that the distance to any set in $\mathcal{G}$ has a closed-form solution, which enables exactly computing the empirical measure of the ϵ-expansion of any set in $\mathcal{G}$.

In addition, it will be easier to control the optimization error (i.e., develop an algorithm that produces tight estimates), if the empirical concentration problem is simpler. For instance, when the underlying perturbation metric is ℓ_∞-norm, solving the empirical concentration problem with respect to the set of half spaces is easier than solving it based on the union of hyperrectangles, since there are more hyperparameters to optimize for the latter problem. Such simplicity further contributes to tighter empirical estimates produced by our algorithm using half spaces for measuring concentration (see Section 3.5 for a detailed argument).

Connection with other Related Works. A line of research studied the inherent tradeoff between robustness and accuracy, which are related to ours. In particular, the work of

[119] showed that for some specific learning problems, achieving robustness and accuracy together is not possible. At the first glance, it might seem that this trade-off contradicts the upper bounds on adversarial robustness that come from concentration of measure. However, there is no contradiction and what is proved there is with regard to a different definition of adversarial examples. The definition of adversarial examples used there could diverge from our definition in some learning problems (see [30]), but they coincide in the cases that the ground-truth function is robust to small perturbations. In addition, they considered adversaries bounded in ℓ_∞-norm with attack radius greater than the distance between the two classes' mean value difference, which is not realistic for classification tasks of interest.

The work of [126] considered a more realistic assumption. They showed that if the underlying data are well-separated, there exists a robust and perfectly accurate classifier, which somewhat suggests that robustness and accuracy can be achieved together in principle. However, this does not contradict our results, since we directly assume the existence of the underlying ground-truth as the concept function in our adversarial risk definition, and study the maximum achievable robustness that can be attained by classifiers with imperfect risk, which implicitly excludes the ground-truth classifier. In addition, the construction of the 'perfectly' robust and accurate classifier presented in [126] assumes the knowledge of the support of the underlying data manifold, which is difficult to test for typical classification tasks; whereas our results do not rely on such assumption. We present a method with provable guarantees for characterizing the concentration function that works for typical classification tasks of interest.

Other related works [129, 99] studied the robustness and accuracy trade-off with respect to specifically learned classifiers. In contrast, our results on intrinsic robustness translated by concentration of measure is different, since it is with respect to the optimally robust classifier that can be *constructed* within the set of imperfect classifiers. Actually *finding*

such an optimal classifier using a learning algorithm might be a more difficult task or even infeasible. We do not consider that problem in this chapter.

3.7 Summary

To understand whether theoretical results showing limits of intrinsic robustness for theoretical distributions apply to robust classification tasks of interest, we developed general methods in this chapter to understand and estimate the concentration function of an unknown distribution. According to our experimental results, we demonstrate that the concentration of measure phenomenon is not the major reason behind vulnerability of the existing classifiers to adversarial examples. In other words, recent impossibility results [49, 44, 80, 108] should not cause us to lose hope in the possibility of finding more robust classifiers. Thus, either there is room for improving the robustness of image classifiers (even with non-zero classification error) or a need for deeper understanding of the reasons for the gap between intrinsic robustness and the actual robustness achieved by robust models, at least for the datasets like the image classification benchmarks used in our experiments. In the next chapter, we are going to study another fundamental cause, i.e. the existence of uncertain inputs, in explaining the adversarial vulnerability of state-of-the-art classification models by adapting our concentration estimation method proposed in Section 3.4.

Chapter 4

Importance of Labels[1]

4.1 Introduction

In this work, we argue that the standard concentration of measure problem, which was studied in all of the aforementioned works, is not sufficient to capture a realistic intrinsic robustness limit for a classification problem. In particular, the standard concentration function (see Definition 3.3) is defined as an inherent property regarding the input metric probability space that does not take account of the underlying label information. We argue that such label information is essential for any supervised learning problem, including adversarially robust classification, so must be incorporated into intrinsic robustness limits.

We introduce a notion of *label uncertainty*, which characterizes the average uncertainty of label assignments for an input region. We then incorporate label uncertainty in the standard concentration measure as an initial step towards a more realistic characterization of intrinsic robustness. Experiments on the CIFAR-10 and CIFAR-10H [95] datasets demonstrate that error regions induced by state-of-the-art classification models all have high label uncertainty, which validates the proposed label uncertainty constrained concentration problem.

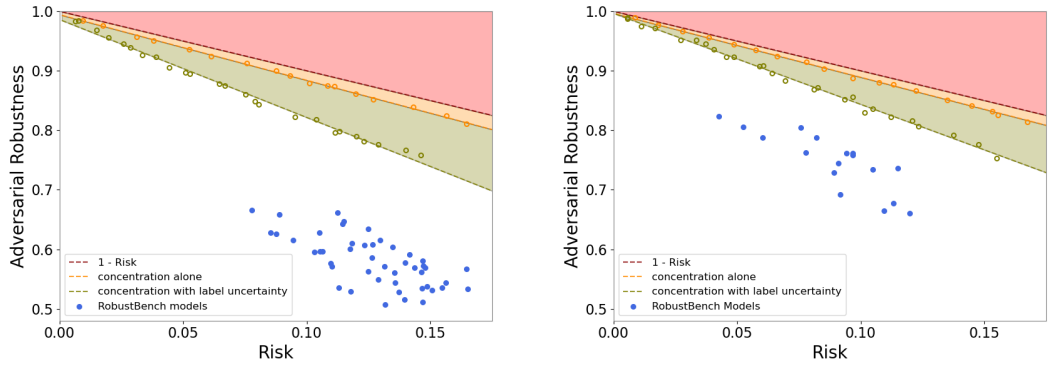
By adapting the standard concentration estimation method in [81], we propose an empirical estimator for the label uncertainty constrained concentration function. We then theoreti-

[1] Xiao Zhang, David Evans, *Understanding Intrinsic Robustness using Label Uncertainty*, in the Tenth International Conference on Learning Representations (ICLR 2022) [133].

cally study the asymptotic behavior of the proposed estimator and provide a corresponding heuristic algorithm for typical perturbation metrics. We demonstrate that our method is able to produce a more accurate characterization of intrinsic robustness limit for benchmark datasets than was possible using prior methods that do not consider labels. Figure 4.1 illustrates the intrinsic robustness estimates resulting from our label uncertainty approach on two CIFAR-10 robust classification tasks. The intrinsic robustness estimates we obtain by incorporating label uncertainty are much lower than prior limits, suggesting that compared with the concentration of measure phenomenon, the existence of uncertain inputs may explain more fundamentally the adversarial vulnerability of state-of-the-art robustly-trained models. In addition, we also provide empirical evidence showing that both the clean and robust accuracies of state-of-the-art robust classification models are largely affected by the label uncertainty of the tested examples, suggesting that adding an abstain option based on label uncertainty is a promising avenue for improving adversarial robustness of deployed machine learning systems.

4.2 Standard Concentration is Insufficient

Concentration without Labels Mischaracterizes Intrinsic Robustness. Despite the appealing relationship between concentration of measure and intrinsic robustness, we argue that solving the standard concentration problem is not enough to capture a meaningful intrinsic limit for adversarially robust classification. The standard concentration of measure problem (3.1), which aims to find the optimal subset that has the smallest ϵ-expansion with regard to the input metric probability space $(\mathcal{X}, \mu, \Delta)$, does not involve the concept function $c(\cdot)$ that determines the underlying class label of each input. Therefore, no matter how we assign the labels to the inputs, the concentration function $h(\mu, \alpha, \epsilon)$ will remain the same for



(a) ℓ_∞ perturbations (b) ℓ_2 perturbations

Figure 4.1: Intrinsic robustness estimates for classification tasks on CIFAR-10 under (a) ℓ_∞ perturbations with $\epsilon = 8/255$ and (b) ℓ_2 perturbations with $\epsilon = 0.5$. Orange dots are intrinsic robustness estimates using the method in [96], which does not consider labels; green dots show the results using our methods that incorporate label uncertainty; blue dots are results achieved by the state-of-the-art adversarially-trained models in RobustBench [21]. Three fundamental causes behind the adversarial vulnerability can be summarized as imperfect risk (red region), concentration of measure (orange region) and existence of uncertain inputs (green region).

the considered metric probability space. In sharp contrast, learning an adversarially-robust classifier depends on the joint distribution of both the inputs and the labels.

Moreover, when the standard concentration function is translated into an intrinsic limit of adversarial robustness, it is defined with respect to the set of imperfect classfiers $\mathcal{F}_\alpha$ (see Theorem 3.5). The only restriction imposed by $\mathcal{F}_\alpha$ is that the classifier (or equivalently, the measure of the corresponding error region) has risk at least α. This fails to consider whether the classifier is learnable or not under the given classification problem. Therefore, the intrinsic robustness limit implied by standard concentration $\overline{\text{AdvRob}}_\epsilon(\mathcal{F}_\alpha)$ could be much higher than $\overline{\text{AdvRob}}_\epsilon(\mathcal{F}_{\text{learn}})$, where $\mathcal{F}_{\text{learn}}$ denotes the set of classifiers that can be produced by some supervised learning method. Hence, it is not surprising that the adversarial robustness attained by state-of-the-art robust training methods for several image benchmarks is much lower than the intrinsic robustness limit implied by standard concentration of measure. In this work, to obtain a more meaningful intrinsic robustness limit we

restrict the search space of the standard concentration problem (3.1) by considering both the underlying class labels and the learnability of the given classification problem.

Gaussian Mixture Model. We further illustrate the insufficiency of standard concentration under a simple Gaussian mixture model. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be the input space and $\mathcal{Y} = \{-1, +1\}$ be the label space. Assume all the inputs are first generated according to a mixture of 2-Gaussian distribution: $\boldsymbol{x} \sim \mu = \frac{1}{2}\mathcal{N}(-\boldsymbol{\theta}, \sigma^2\mathbf{I}_n) + \frac{1}{2}\mathcal{N}(\boldsymbol{\theta}, \sigma^2\mathbf{I}_n)$, then labeled by a concept function $c(\boldsymbol{x}) = \text{sgn}(\boldsymbol{\theta}^\top \boldsymbol{x})$, where $\boldsymbol{\theta} \in \mathbb{R}^n$ and $\sigma \in \mathbb{R}$ are given parameters (this concept function is also the Bayes optimal classifier, which best separates the two Gaussian clusters). Theorem 4.1 characterizes the optimal solution to the standard concentration problem under this assumed model.

Theorem 4.1. *Consider the above Gaussian mixture model with ℓ_2 perturbation metric. The optimal solution to the standard concentration problem* (3.1) *is a halfspace, either*

$$\mathcal{H}_- = \{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{\theta}^\top \boldsymbol{x} + b \cdot \|\boldsymbol{\theta}\|_2 \leq 0\} \quad \textit{or} \quad \mathcal{H}_+ = \{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{\theta}^\top \boldsymbol{x} - b \cdot \|\boldsymbol{\theta}\|_2 \geq 0\},$$

where b is a parameter depending on α and $\boldsymbol{\theta}$ such that $\mu(\mathcal{H}_-) = \mu(\mathcal{H}_+) = \alpha$.

Proof of Theorem 4.1. Let μ_- be the probability measure for $\mathcal{N}(-\boldsymbol{\theta}, \sigma^2\mathbf{I}_n)$ and μ_+ be the probability measure for $\mathcal{N}(\boldsymbol{\theta}, \sigma^2\mathbf{I}_n)$, then by definition, we have $\mu = \mu_-/2 + \mu_+/2$. Consider the optimal subset $\mathcal{E}^* = \text{argmin}_{\mathcal{E} \in \text{Pow}(\mathcal{X})}\{\mu_\epsilon(\mathcal{E}) : \mu(\mathcal{E}) \geq \alpha\}$.

Note that the standard concentration function $h(\mu, \alpha, \epsilon)$ is monotonically increasing with respect to α, thus $\mu(\mathcal{E}^*) = \alpha$ holds for any continuous μ. Let $\alpha_- = \mu_-(\mathcal{E}^*)$ and $\alpha_+ = \mu_+(\mathcal{E}^*)$. According to the Gaussian Isoperimetric Inequality (see Lemma 3.4), it holds for

any $\epsilon \geq 0$ that

$$\mu(\mathcal{E}_\epsilon^*) = \frac{1}{2}\mu_-(\mathcal{E}_\epsilon^*) + \frac{1}{2}\mu_+(\mathcal{E}_\epsilon^*) \geq \frac{1}{2}\Phi(\Phi^{-1}(\alpha_-) + \epsilon) + \frac{1}{2}\Phi(\Phi^{-1}(\alpha_+) + \epsilon). \tag{4.1}$$

Note that the equality of (4.1) can be achieved if and only if $\mathcal{E}^*$ is a half space.

Next, we show that there always exists a half space $\mathcal{H} \in \text{Pow}(\mathcal{X})$ such that $\mu_-(\mathcal{H}) = \alpha_-$ and $\mu_+(\mathcal{H}) = \alpha_+$. Let $f_-(\cdot)$, $f_+(\cdot)$ be the PDFs of μ_- and μ_+ respectively. For any $\boldsymbol{x} \in \mathcal{X}$, $f_-(\boldsymbol{x})$ and $f_+(\boldsymbol{x})$ are always positive, thus we have

$$\frac{f_+(\boldsymbol{x})}{f_-(\boldsymbol{x})} = \frac{\exp\{-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\theta})^\top(\boldsymbol{x}-\boldsymbol{\theta})\}}{\exp\{-\frac{1}{2\sigma^2}(\boldsymbol{x}+\boldsymbol{\theta})^\top(\boldsymbol{x}+\boldsymbol{\theta})\}} = \exp\left(\frac{2\boldsymbol{\theta}^\top\boldsymbol{x}}{\sigma^2}\right).$$

This implies that the ratio of $f_+(\boldsymbol{x})/f_-(\boldsymbol{x})$ is monotonically increasing with respect to $\boldsymbol{\theta}^\top\boldsymbol{x}$.

Consider the following extreme half space $\mathcal{H}_- = \{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{\theta}^\top\boldsymbol{x} + b \cdot \|\boldsymbol{\theta}\|_2 \leq 0\}$ such that $\mu(\mathcal{H}_-) = \alpha$. We are going to prove $\mu_-(\mathcal{H}_-) \geq \mu_-(\mathcal{E}^*) = \alpha_-$ and $\mu_+(\mathcal{H}_-) \leq \mu_+(\mathcal{E}^*) = \alpha_+$. Consider the sets $\mathcal{E}^* \cap (\mathcal{H}_-)^\complement$ and $(\mathcal{E}^*)^\complement \cap \mathcal{H}_-$, we have

$$\frac{\mu_+\big(\mathcal{E}^* \cap (\mathcal{H}_-)^\complement\big)}{\mu_-\big(\mathcal{E}^* \cap (\mathcal{H}_-)^\complement\big)} \geq \inf_{\boldsymbol{x} \in \mathcal{E}^* \cap (\mathcal{H}_-)^\complement} \exp\left(\frac{2\boldsymbol{\theta}^\top\boldsymbol{x}}{\sigma^2}\right) \geq \sup_{\boldsymbol{x} \in (\mathcal{E}^*)^\complement \cap \mathcal{H}_-} \left(\frac{2\boldsymbol{\theta}^\top\boldsymbol{x}}{\sigma^2}\right) \geq \frac{\mu_+\big((\mathcal{E}^*)^\complement \cap \mathcal{H}_-\big)}{\mu_-\big((\mathcal{E}^*)^\complement \cap \mathcal{H}_-\big)}. \tag{4.2}$$

Note that we also have

$$\mu_+\big(\mathcal{E}^* \cap (\mathcal{H}_-)^\complement\big) + \mu_-\big(\mathcal{E}^* \cap (\mathcal{H}_-)^\complement = \mu_+\big((\mathcal{E}^*)^\complement \cap \mathcal{H}_-\big) + \mu_-\big((\mathcal{E}^*)^\complement \cap \mathcal{H}_-\big). \tag{4.3}$$

Thus, combining (4.2) and (4.3), we have

$$\mu_+\big(\mathcal{E}^* \cap (\mathcal{H}_-)^\complement\big) \geq \mu_+\big((\mathcal{E}^*)^\complement \cap \mathcal{H}_-\big) \quad \text{and} \quad \mu_-\big(\mathcal{E}^* \cap (\mathcal{H}_-)^\complement\big) \leq \mu_-\big((\mathcal{E}^*)^\complement \cap \mathcal{H}_-\big),$$

Adding the term $\mu_+\big(\mathcal{E}^* \cap \mathcal{H}_-\big)$ or $\mu_-\big(\mathcal{E}^* \cap \mathcal{H}_-\big)$ on both sides, we further have

$$\mu_+(\mathcal{H}_-) \leq \mu_+(\mathcal{E}^*) = \alpha_+ \quad \text{and} \quad \mu_-(\mathcal{H}_-) \geq \mu_-(\mathcal{E}^*) = \alpha_-.$$

On the other hand, consider the half space $\mathcal{H}_+ = \{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{\theta}^\top \boldsymbol{x} - b \cdot \|\boldsymbol{\theta}\|_2 \geq 0\}$ such that $\mu(\mathcal{H}_+) = \alpha$. Based on a similar technique, we can prove

$$\mu_-(\mathcal{H}_+) \geq \mu_+(\mathcal{E}^*) = \alpha_+ \quad \text{and} \quad \mu_-(\mathcal{H}_+) \leq \mu_-(\mathcal{E}^*) = \alpha_-.$$

In addition, let $\mathcal{H} = \{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{w}^\top \boldsymbol{x} + b \leq 0\}$ be any half space such that $\mu(\mathcal{H}) = \alpha$. Since both μ_+ and μ_- are continuous, as we rotate the half space (i.e., gradually increase the value of $\boldsymbol{w}^\top \boldsymbol{\theta}$), $\mu_-(\mathcal{H})$ and $\mu_+(\mathcal{H})$ will also change continuously. Therefore, it is guaranteed that there exists a half space $\mathcal{H} \in \text{Pow}(\mathcal{X})$ such that $\mu_-(\mathcal{H}) = \alpha_-$ and $\mu_+(\mathcal{H}) = \alpha_+$. This further implies that the lower bound of (4.1) can be always be achieved.

Finally, since we have proved the optimal subset has to be a half space, the remaining task is to solve the following optimization problem:

$$\begin{aligned} \min_{\mathcal{H} \in \text{Pow}(\mathcal{X})} \; & \frac{1}{2}\Phi\big(\Phi^{-1}\big(\mu_-(\mathcal{H})\big) + \epsilon\big) + \frac{1}{2}\Phi\big(\Phi^{-1}\big(\mu_+(\mathcal{H})\big) + \epsilon\big) \\ \text{s.t. } & \mathcal{H} = \{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{w}^\top \boldsymbol{x} + b \leq 0\} \quad \text{and} \quad \mu(\mathcal{H}) = \alpha. \end{aligned} \tag{4.4}$$

Construct function $g(u) = \Phi\big(\Phi^{-1}(u) + \epsilon\big) + \Phi\big(\Phi^{-1}(2\alpha - u) + \epsilon\big)$, where $u \in [0, 2\alpha]$. Based on the derivative of inverse function formula, we compute the derivative of g with respect

to u as follows

$$\begin{aligned}
\frac{\mathrm{d}g(u)}{\mathrm{d}u} &= \frac{1}{\sqrt{2\pi}}\exp\Big\{-\frac{(\Phi^{-1}(u)+\epsilon)^2}{2}\Big\}\cdot\frac{\mathrm{d}\Phi^{-1}(u)}{\mathrm{d}u} \\
&\qquad + \frac{1}{\sqrt{2\pi}}\exp\Big\{-\frac{(\Phi^{-1}(2\alpha-u)+\epsilon)^2}{2}\Big\}\cdot\frac{\mathrm{d}\Phi^{-1}(2\alpha-u)}{\mathrm{d}u} \\
&= \exp\Big\{-\frac{(\Phi^{-1}(u)+\epsilon)^2}{2}\Big\}\cdot\exp\Big\{\frac{(\Phi^{-1}(u))^2}{2}\Big\} \\
&\qquad - \exp\Big\{-\frac{(\Phi^{-1}(2\alpha-u)+\epsilon)^2}{2}\Big\}\cdot\exp\Big\{\frac{(\Phi^{-1}(2\alpha-u))^2}{2}\Big\} \\
&= \exp(-\epsilon^2/2)\cdot\Big[\exp\big(-\epsilon\Phi^{-1}(u)\big)-\exp\big(-\epsilon\Phi^{-1}(2\alpha-u)\big)\Big].
\end{aligned}$$

Noticing the term $\exp(-\epsilon\Phi^{-1}(u))$ is monotonically decreasing with respect to u, we then know that $g(u)$ is monotonically increasing in $[0,\alpha]$ and monotonically decreasing in $[\alpha,2\alpha]$. Therefore, this suggests that the optimal solution to (4.4) is achieved when $\mu_-(\mathcal{H})$ reaches its maximum or its minimum. According to the previous argument regarding the range of α_- and α_+, we can immediately prove the optimality results of Theorem 4.1. □

Remark 4.2. Theorem 4.1 suggests that for the Gaussian mixture model, the optimal subset achieving the smallest ϵ-expansion under ℓ_2-norm distance metric is a halfspace $\mathcal{H}$, which is far away from the boundary between the two Gaussian classes for small α. When translated into the intrinsic robustness problem, the corresponding optimal classifier f has to be constructed by treating $\mathcal{H}$ as the only error region, or more precisely $f(\boldsymbol{x}) = c(\boldsymbol{x})$ if $\boldsymbol{x} \notin \mathcal{H}$; $f(\boldsymbol{x}) \neq c(\boldsymbol{x})$ otherwise. This optimally constructed classifier f, however, does not match our intuition of what a predictive classifier would do under the considered Gaussian mixture model. In particular, since all the inputs in $\mathcal{H}$ and their neighbours share the same class label and are also far away from the boundary, examples that fall into $\mathcal{H}$ should be easily classified correctly using simple decision rule, such as k-nearest neighbour or maximum margin, whereas examples that are close to the boundary should be more likely to

be misclassified as errors by supervisedly-learned classifiers. This confirms our claim that standard concentration is not sufficient for capturing a meaningful intrinsic robustness limit.

4.3 Incorporating Labels in Intrinsic Robustness

In this section, we first propose a new concentration estimation framework by imposing a constraint based on label uncertainty (Definition 4.3) on the search space with respect to the standard problem (3.1). Then, we explain why this yields a more realistic intrinsic robustness limit.

Let $(\mathcal{X}, \mu)$ be the input probability space and $\mathcal{Y} = \{1, 2, \ldots, k\}$ be the set of labels. $\eta : \mathcal{X} \to [0,1]^k$ is said to capture the *full label distribution* [47, 45], 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 $\sum_{y \in [k]} [\eta(\boldsymbol{x})]_y = 1$ holds for any $\boldsymbol{x} \in \mathcal{X}$. For classification tasks that rely on human labeling, one can approximate the label distribution for any input by collecting human labels from multiple human annotators. Our experiments use the CIFAR-10H dataset that did this for the CIFAR-10 test images [95].

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

Definition 4.3 (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 $c : \mathcal{X} \to \mathcal{Y}$ is a concept function that assigns each input $\boldsymbol{x}$ a label $y \in \mathcal{Y}$. Assume $\eta : \mathcal{X} \to [0,1]^k$ is the underlying label distribution function, where $[\eta(\boldsymbol{x})]_y$ represents the description degree of y to $\boldsymbol{x}$. For any subset $\mathcal{E} \in \text{Pow}(\mathcal{X})$ with measure $\mu(\mathcal{E}) > 0$, the *label uncertainty* (LU) of $\mathcal{E}$ with respect

to $(\mathcal{X}, \mu)$, $c(\cdot)$ and $\eta(\cdot)$ is defined as:

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

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 $\{\boldsymbol{x}, c(\boldsymbol{x})\}$. The range of label uncertainty is $[0, 2]$. 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; label uncertainty of 2 means the input is mislabeled and there is a different label that represents the ground-truth label. Let $(\mathcal{X}, \mu, \Delta)$ be the underlying input metric probability space. Based on the notion of label uncertainty, we study the following constrained concentration problem:

$$\underset{\mathcal{E} \in \mathrm{Pow}(\mathcal{X})}{\text{minimize}} \ \mu(\mathcal{E}_\epsilon) \quad \text{subject to} \quad \mu(\mathcal{E}) \geq \alpha \ \text{ and } \ \mathrm{LU}(\mathcal{E}; \mu, c, \eta) \geq \gamma, \tag{4.5}$$

where $\gamma \in [0, 2]$ is a constant. When γ is set as zero, (4.5) simplifies to the standard concentration of measure problem. We set the value of γ to roughly represent the label uncertainty of the error region of state-of-the-art classifiers for the given classification problem.

Theorem 4.4 shows how (4.5) captures the intrinsic robustness limit with respect to the set of imperfect classifiers whose error region label uncertainty is at least γ.

Theorem 4.4. *Define* $\mathcal{F}_{\alpha,\gamma} = \{f : \mathrm{Risk}(f) \geq \alpha, \ \mathrm{LU}(\mathcal{E}_f; \mu, c, \eta) \geq \gamma\}$*, where* $\alpha \in (0, 1)$*,* $\gamma \in (0, 2)$ *and* $\mathcal{E}_f = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \neq c(\boldsymbol{x})\}$ *is the error region of* f*. For any* $\epsilon \geq 0$*, it*

holds that

$$\inf_{\mathcal{E}\in\text{Pow}(\mathcal{X})}\{\mu(\mathcal{E}_\epsilon) : \mu(\mathcal{E}) \geq \alpha,\ \text{LU}(\mathcal{E};\mu,c,\eta) \geq \gamma\} = 1 - \overline{\text{AdvRob}}_\epsilon(\mathcal{F}_{\alpha,\gamma}).$$

Proof of Theorem 4.4. Let $\mathcal{E}^*$ be the optimal solution to (4.5), then $\mu(\mathcal{E}_\epsilon^*)$ is the optimal value of (4.5). We will show $1 - \overline{\text{AdvRob}}_\epsilon(\mathcal{F}_{\alpha,\gamma}) = \mu(\mathcal{E}_\epsilon^*)$ by proving both directions.

First, we prove $1 - \overline{\text{AdvRob}}_\epsilon(\mathcal{F}_{\alpha,\gamma}) \geq \mu(\mathcal{E}_\epsilon^*)$. Let f be any classifier within $\mathcal{F}_{\alpha,\gamma}$, and $\mathcal{E}(f)$ be the corresponding error region of f. According to Definition 3.1, we have

$$\text{Risk}(f) = \mu(\mathcal{E}(f)) \ \text{ and } \ \text{AdvRisk}_\epsilon(f) = \mu(\mathcal{E}_\epsilon(f)),$$

where $\mathcal{E}_\epsilon(f)$ represents the ϵ-expansion of $\mathcal{E}(f)$. Since $f \in \mathcal{F}_{\alpha,\gamma}$, we have

$$\text{Risk}(f) = \mu(\mathcal{E}(f)) \geq \alpha \ \text{ and } \ \text{LU}(\mathcal{E}(f);\mu,c,\eta) \geq \gamma.$$

Thus, by (4.5), we obtain that

$$1 - \text{AdvRob}_\epsilon(f) = \text{AdvRisk}_\epsilon(f) = \mu(\mathcal{E}_\epsilon(f)) \geq \mu(\mathcal{E}_\epsilon^*).$$

By taking the infimum over f over $\mathcal{F}_{\alpha,\gamma}$, we have $1 - \overline{\text{AdvRob}}_\epsilon(\mathcal{F}_\alpha) \geq \mu(\mathcal{E}_\epsilon^*)$.

Next, we show that $1 - \overline{\text{AdvRob}}_\epsilon(\mathcal{F}_{\alpha,\gamma}) \leq \mu(\mathcal{E}_\epsilon^*)$. We construct a classifier f^* such that

$$f^*(\boldsymbol{x}) = c(\boldsymbol{x}) \ \text{ if } \ \boldsymbol{x} \notin \mathcal{E}^*;\ f^*(\boldsymbol{x}) \neq c(\boldsymbol{x}) \ \text{ otherwise.}$$

Note that by construction, $\mathcal{E}^*$ corresponds to the error region of f^*. Thus according to the

definitions of risk and adversarial risk, we know

$$\text{Risk}(f^*) = \mu(\mathcal{E}^*) \geq \alpha \text{ and } \text{AdvRisk}_\epsilon(f^*) = \mu(\mathcal{E}^*_\epsilon).$$

Since $\text{LU}(\mathcal{E}^*; \mu, c, \eta) \geq \gamma$, the error region label uncertainty of f^* is at least γ. Thus, by definition of intrinsic robustness, we know $1 - \overline{\text{AdvRob}}_\epsilon(\mathcal{F}_{\alpha,\gamma}) \leq \text{AdvRisk}_\epsilon(f^*) = \mu(\mathcal{E}^*_\epsilon)$.

Finally, putting pieces together, we complete the proof. □

Compared with standard concentration, (4.5) aims to search for the least expansive subset with respect to input regions with high label uncertainty. According to Theorem 4.4, the translated intrinsic robustness limit is defined with respect to $\mathcal{F}_{\alpha,\gamma}$ and is guaranteed to be no greater than $\overline{\text{AdvRob}}_\epsilon(\mathcal{F}_\alpha)$. Although both $\overline{\text{AdvRob}}_\epsilon(\mathcal{F}_\alpha)$ and $\overline{\text{AdvRob}}_\epsilon(\mathcal{F}_{\alpha,\gamma})$ can serve as valid robustness upper bounds for any $f \in \mathcal{F}_{\alpha,\gamma}$, the latter one would be able to capture a more meaningful intrinsic robustness limit, since state-of-the-art classifiers are expected to more frequently misclassify inputs with large label uncertainty, as there is more discrepancy between their assigned labels and the underlying label distribution (Section 4.5 provides supporting empirical evidence for this on CIFAR-10).

Need for Soft Labels. The proposed approach requires label uncertainty information for training examples. The CIFAR-10H dataset provided soft labels from humans that enabled our experiments, but typical machine learning datasets do not provide such information. Below, we discuss possible avenues to estimating label uncertainty when human soft labels are not available and are too expensive to acquire. A potential solution is to estimate the set of examples with high label uncertainty using the predicted probabilities of a classification model. Confident learning [88, 76, 60, 91, 90] provides a systematic method to identify label errors in a dataset based on this idea. If the estimated label errors match the examples with high human label uncertainty, then we can directly extend our framework by leverag-

ing the estimated error set. Our experiments on CIFAR-10 (see Section 4.5.4), however, suggest that there is a misalignment between human recognized errors and errors produced by confident learning. The existence of such misalignment further suggests that one should be cautious when combining the estimated set of label errors into our framework. As the field of confident learning advances to produce a more accurate estimator of label error set, it would serve as a good alternative solution for applying our framework to the setting where human label information is not accessible.

4.4 Measuring Concentration with Label Uncertainty

Directly solving (4.5) requires the knowledge of the underlying input distribution μ and the ground-truth label distribution function $\eta(\cdot)$, which are usually not available for classification problems. Thus, we consider the following empirical counterpart of (4.5):

$$\underset{\mathcal{E}\in\mathcal{G}}{\text{minimize}}\ \hat{\mu}_{\mathcal{S}}(\mathcal{E}_\epsilon) \quad \text{subject to} \quad \hat{\mu}_{\mathcal{S}}(\mathcal{E}) \geq \alpha \ \text{ and } \ \mathrm{LU}(\mathcal{E}; \hat{\mu}_{\mathcal{S}}, c, \hat{\eta}) \geq \gamma, \tag{4.6}$$

where the search space is restricted to some specific collection of subsets $\mathcal{G} \subseteq \mathrm{Pow}(\mathcal{X})$, μ is replaced by the empirical distribution $\hat{\mu}_{\mathcal{S}}$ with respect to a set of inputs sampled from μ, and the empirical label distribution $\hat{\eta}(\boldsymbol{x})$ is considered as an empirical replacement of $\eta(\boldsymbol{x})$ for any given input $\boldsymbol{x} \in \mathcal{S}$.

The following theorem characterizes a generalization bound regarding the proposed label uncertainty estimate. It shows that if $\mathcal{G}$ is not too complex and $\hat{\eta}$ is close to the ground-truth label distribution function η, the empirical estimate of label uncertainty $\mathrm{LU}(\mathcal{E}; \hat{\mu}_{\mathcal{S}}, c, \hat{\eta})$ is guaranteed to be close to the actual label uncertainty $\mathrm{LU}(\mathcal{E}; \mu, c, \eta)$.

Theorem 4.5 (Generalization of Label Uncertainty)**.** *Let $(\mathcal{X}, \mu)$ be a probability space and*

$\mathcal{G} \subseteq \text{Pow}(\mathcal{X})$ be a collection of subsets of $\mathcal{X}$. Assume $\phi : \mathbb{N} \times \mathbb{R} \to [0, 1]$ is a complexity penalty for $\mathcal{G}$. If $\hat{\eta}(\cdot)$ is close to $\eta(\cdot)$ in L^1-norm with respect to μ, i.e. $\int_{\mathcal{X}} \|\eta(\boldsymbol{x}) - \hat{\eta}(\boldsymbol{x})\|_1 d\mu \leq \delta_\eta$, where $\delta_\eta \in (0, 1)$ is a small constant, then for any $\alpha, \delta \in (0, 1)$ such that $\delta < \alpha$, we have

$$\Pr_{\mathcal{S} \leftarrow \mu^m} \left[\exists\, \mathcal{E} \in \mathcal{G} \text{ and } \mu(\mathcal{E}) \geq \alpha : \left| \text{LU}(\mathcal{E}; \mu, c, \eta) - \text{LU}(\mathcal{E}; \hat{\mu}_{\mathcal{S}}, c, \hat{\eta}) \right| \leq \frac{4\delta + \delta_\eta}{\alpha - \delta} \right] \leq \phi(m, \delta).$$

Proof of Theorem 4.5. Let $\text{lu}(\boldsymbol{x}; c, \eta) = 1 - \left[\eta(\boldsymbol{x})\right]_{c(\boldsymbol{x})} + \max_{y' \neq c(\boldsymbol{x})} \left[\eta(\boldsymbol{x})\right]_{y'}$ be the label uncertainty of a given input $\boldsymbol{x}$ with respect to $c(\cdot)$ and $\eta(\cdot)$. Let $\mathcal{E}$ be a subset in $\mathcal{G}$ such that $\mu(\mathcal{E}) \geq \alpha$ and $|\mu(\mathcal{E}) - \hat{\mu}(\mathcal{E})| \leq \delta$, where δ is a constant much smaller than α. Then according to Definition 4.3, we can decompose the estimation error of label uncertainty as:

$$\begin{aligned}
\text{LU}(\mathcal{E}; \mu, c, \eta) - \text{LU}(\mathcal{E}; \hat{\mu}_{\mathcal{S}}, c, \hat{\eta}) &= \frac{1}{\mu(\mathcal{E})} \int_{\mathcal{E}} \text{lu}(\boldsymbol{x}; c, \eta)\, d\mu - \frac{1}{\hat{\mu}_{\mathcal{S}}(\mathcal{E})} \int_{\mathcal{E}} \text{lu}(\boldsymbol{x}; c, \hat{\eta})\, d\hat{\mu}_{\mathcal{S}} \\
&= \underbrace{\left(\frac{1}{\mu(\mathcal{E})} - \frac{1}{\hat{\mu}_{\mathcal{S}}(\mathcal{E})} \right) \cdot \int_{\mathcal{E}} \text{lu}(\boldsymbol{x}; c, \eta)\, d\mu}_{I_1} \\
&\quad + \underbrace{\frac{1}{\hat{\mu}_{\mathcal{S}}(\mathcal{E})} \int_{\mathcal{E}} \left[\text{lu}(\boldsymbol{x}; c, \eta) - \text{lu}(\boldsymbol{x}; c, \hat{\eta}) \right] d\mu}_{I_2} \\
&\quad + \underbrace{\frac{1}{\hat{\mu}_{\mathcal{S}}(\mathcal{E})} \left(\int_{\mathcal{E}} \text{lu}(\boldsymbol{x}; c, \hat{\eta})\, d\mu - \int_{\mathcal{E}} \text{lu}(\boldsymbol{x}; c, \hat{\eta})\, d\hat{\mu}_{\mathcal{S}} \right)}_{I_3}.
\end{aligned}$$

Next, we upper bound the absolute value of the three components, respectively.

Consider the first term I_1. Note that $0 \leq \text{lu}(\boldsymbol{x}; c, \eta) \leq 2$ for any $\boldsymbol{x} \in \mathcal{X}$, thus we have $|\int_{\mathcal{E}} \text{lu}(\boldsymbol{x}; c, \eta)\, d\mu| \leq 2\mu(\mathcal{E})$. Therefore, we have

$$|I_1| \leq \left| \frac{1}{\mu(\mathcal{E})} - \frac{1}{\hat{\mu}_{\mathcal{S}}(\mathcal{E})} \right| \cdot 2\mu(\mathcal{E}) \leq \frac{2}{\hat{\mu}_{\mathcal{S}}(\mathcal{E})} \cdot |\mu(\mathcal{E}) - \hat{\mu}_{\mathcal{S}}(\mathcal{E})|.$$

As for the second term I_2, the following inequality holds for any $\boldsymbol{x} \in \mathcal{X}$

$$\begin{aligned}|\text{lu}(\boldsymbol{x}; c, \eta) - \text{lu}(\boldsymbol{x}; c, \hat{\eta})| &\leq \left| \left[\eta(\boldsymbol{x}) - \hat{\eta}(\boldsymbol{x})\right]_{c(\boldsymbol{x})} \right| + \left| \max_{y' \neq c(\boldsymbol{x})} \left[\eta(\boldsymbol{x})\right]_{y'} - \max_{y' \neq c(\boldsymbol{x})} \left[\hat{\eta}(\boldsymbol{x})\right]_{y'} \right| \\ &\leq \left\| \eta(\boldsymbol{x}) - \hat{\eta}(\boldsymbol{x}) \right\|_1,\end{aligned}$$

where the second inequality holds because $|\max_i a_i - \max_i b_i| \leq \max_i |a_i - b_i|$ for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$. Therefore, we can upper bound $|I_2|$ by

$$|I_2| \leq \frac{1}{\hat{\mu}_{\mathcal{S}}(\mathcal{E})} \int_{\mathcal{E}} \left\| \eta(\boldsymbol{x}) - \hat{\eta}(\boldsymbol{x}) \right\|_1 d\mu \leq \frac{1}{\hat{\mu}_{\mathcal{S}}(\mathcal{E})} \int_{\mathcal{X}} \left\| \eta(\boldsymbol{x}) - \hat{\eta}(\boldsymbol{x}) \right\|_1 d\mu \leq \frac{\delta_\eta}{\hat{\mu}_{\mathcal{S}}(\mathcal{E})}.$$

For the last term I_3, since $0 \leq \text{lu}(\boldsymbol{x}; c, \eta) \leq 2$ holds for any $\boldsymbol{x} \in \mathcal{X}$, we have

$$|I_3| \leq \frac{2}{\hat{\mu}_{\mathcal{S}}(\mathcal{E})} \cdot \left| \mu(\mathcal{E}) - \hat{\mu}_{\mathcal{S}}(\mathcal{E}) \right|.$$

Finally, putting pieces together, we have

$$|\text{LU}(\mathcal{E}; \mu, c, \eta) - \text{LU}(\mathcal{E}; \hat{\mu}_{\mathcal{S}}, c, \hat{\eta})| \leq \frac{4}{\hat{\mu}_{\mathcal{S}}(\mathcal{E})} \cdot \left| \mu(\mathcal{E}) - \hat{\mu}_{\mathcal{S}}(\mathcal{E}) \right| + \frac{\delta_\eta}{\hat{\mu}_{\mathcal{S}}(\mathcal{E})} \leq \frac{4\delta + \delta_\eta}{\alpha - \delta},$$

provided $\mu(\mathcal{E}) \geq \alpha$ and $|\mu(\mathcal{E}) - \hat{\mu}_{\mathcal{S}}(\mathcal{E})| \leq \delta$. Making use of the definition of complexity penalty for $\mathcal{G}$ completes the proof of Theorem 4.5. □

Theorem 4.5 implies the generalization of concentration under label uncertainty constraints. The following theorem, inspired by Theorem 3.13, shows that if we choose $\mathcal{G}$ and the collection of its ϵ-expansions, $\mathcal{G}_\epsilon = \{\mathcal{E}_\epsilon : \mathcal{E} \in \mathcal{G}\}$ in a careful way that both of their complexities are small, then with high probability, the empirical label uncertainty constrained concentration will be close to the actual concentration when the search space is restricted to $\mathcal{G}$.

Theorem 4.6 (Generalization of Concentration)**.** *Let $(\mathcal{X}, \mu, \Delta)$ be a metric probability*

space and $\mathcal{G} \subseteq \text{Pow}(\mathcal{X})$. *Define the generalized concentration function under label uncertainty constraints as* $h(\mu, c, \eta, \alpha, \gamma, \epsilon, \mathcal{G}) = \inf_{\mathcal{E}\in\mathcal{G}}\{\mu(\mathcal{E}_\epsilon) : \mu(\mathcal{E}) \geq \alpha, \text{LU}(\mathcal{E}; \mu, c, \eta) \geq \gamma\}$. *Then, under the same setting of Theorem 4.5, for any* $\gamma, \epsilon \in [0, 1]$, $\alpha \in (0, 1]$ *and* $\delta \in (0, \alpha/2)$, *we have*

$$\Pr_{\mathcal{S}\leftarrow\mu^m}\big[h(\mu, c, \eta, \alpha - \delta, \gamma - \delta', \epsilon, \mathcal{G}) - \delta \leq h(\hat{\mu}_\mathcal{S}, c, \hat{\eta}, \alpha, \gamma, \epsilon, \mathcal{G}) \\ \leq h(\mu, c, \eta, \alpha + \delta, \gamma + \delta', \epsilon, \mathcal{G}) + \delta\big] \geq 1 - 6\phi(m, \delta) - 2\phi_\epsilon(m, \delta),$$

where $\delta' = (4\delta + \delta_\eta)/(\alpha - 2\delta)$ *and* ϕ_ϵ *is the complexity penalty for* $\mathcal{G}_\epsilon$.

Proof of Theorem 4.6. Our proof is inspired by the techniques used in proving Theorem 3.13. First, we introduce some notations. Let $h(\mu, c, \eta, \alpha, \gamma, \epsilon, \mathcal{G})$ be the optimal value and $g(\mu, c, \eta, \alpha, \gamma, \epsilon, \mathcal{G})$ be the optimal solution with respect to the following generalized concentration of measure problem with label uncertainty constraint:

$$\underset{\mathcal{E}\in\mathcal{G}}{\text{minimize}}\ \mu(\mathcal{E}_\epsilon) \quad \text{subject to} \quad \mu(\mathcal{E}) \geq \alpha \ \text{ and } \ \text{LU}(\mathcal{E}; \mu, c, \eta) \geq \gamma. \tag{4.7}$$

Note that the difference between (4.7) and (4.5) is that the feasible set of $\mathcal{E}$ is restricted to some collection of subsets $\mathcal{G} \subseteq \text{Pow}(\mathcal{X})$. Correspondingly, we let $h(\hat{\mu}_\mathcal{S}, c, \hat{\eta}, \alpha, \gamma, \epsilon, \mathcal{G})$ and $g(\hat{\mu}_\mathcal{S}, c, \hat{\eta}, \alpha, \gamma, \epsilon, \mathcal{G})$ be the optimal value and optimal solution with respect to the empirical optimization problem (4.6).

Let $\mathcal{E} = g(\mu, c, \eta, \alpha + \delta, \gamma + \delta', \epsilon, \mathcal{G})$ and $\hat{\mathcal{E}} = g(\hat{\mu}_\mathcal{S}, c, \hat{\eta}, \alpha, \gamma, \epsilon, \mathcal{G})$, where δ' will be specified later. Note that when these optimal sets do not exist, we can select a set for which the expansion is arbitrarily close to the optimum, then every step of the proof will apply to

this variant. According to the definition of complexity penalty, we have

$$\Pr_{\mathcal{S}\leftarrow\mu^m}\left[|\hat{\mu}_{\mathcal{S}}(\hat{\mathcal{E}}) - \mu(\hat{\mathcal{E}})| \geq \delta\right] \leq \phi(m,\delta). \tag{4.8}$$

Since $\hat{\mu}_{\mathcal{S}}(\hat{\mathcal{E}}) \geq \alpha$ by definition, (4.8) implies that

$$\Pr_{\mathcal{S}\leftarrow\mu^m}\left[\mu(\hat{\mathcal{E}}) \leq \alpha - \delta\right] \leq \phi(m,\delta). \tag{4.9}$$

In addition, according to Theorem 4.5, for any $\delta \in (0, \alpha/2)$, we have

$$\Pr_{\mathcal{S}\leftarrow\mu^m}\left[\left|\mathrm{LU}(\hat{\mathcal{E}};\mu,c,\eta) - \mathrm{LU}(\hat{\mathcal{E}};\hat{\mu}_{\mathcal{S}},c,\hat{\eta})\right| \leq \frac{4\delta+\delta_\eta}{\alpha-2\delta}\right] \leq 2\phi(m,\delta), \tag{4.10}$$

where the inequality holds because of (4.9) and the union bound. Since $\mathrm{LU}(\hat{\mathcal{E}};\hat{\mu}_{\mathcal{S}},c,\hat{\eta}) \geq \gamma$ by definition, (4.10) implies that

$$\Pr_{\mathcal{S}\leftarrow\mu^m}\left[\mathrm{LU}(\hat{\mathcal{E}};\mu,c,\eta) \leq \gamma - \frac{4\delta+\delta_\eta}{\alpha-2\delta}\right] \leq 2\phi(m,\delta). \tag{4.11}$$

Based on the definition of the concentration function h, combining (4.9) and (4.11) and making use of the union bound, we have

$$\Pr_{\mathcal{S}\leftarrow\mu^m}\left[\mu(\hat{\mathcal{E}}_\epsilon) \leq h(\mu,c,\eta,\alpha-\delta,\gamma-\delta',\epsilon,\mathcal{G})\right] \leq 3\phi(m,\delta), \tag{4.12}$$

where we set $\delta' = \frac{4\delta+\delta_\eta}{\alpha-2\delta}$. Note that according to the definition of ϕ_ϵ, we have

$$\Pr_{\mathcal{S}\leftarrow\mu^m}\left[|\mu(\hat{\mathcal{E}}_\epsilon) - \mu_{\mathcal{S}}(\hat{\mathcal{E}}_\epsilon)| \leq \delta\right] \leq \phi_\epsilon(m,\delta), \tag{4.13}$$

thus combining (4.12) and (4.13) by union bound, we have

$$\Pr_{\mathcal{S}\leftarrow\mu^m}\left[\hat{\mu}_{\mathcal{S}}(\hat{\mathcal{E}}_\epsilon) \leq h(\mu, c, \eta, \alpha-\delta, \gamma-\delta', \epsilon, \mathcal{G}) - \delta\right] \leq 3\phi(m,\delta) + \phi_\epsilon(m,\delta). \tag{4.14}$$

This completes the proof of one-sided inequality of Theorem 4.6. The other side of Theorem 4.6 can be proved using the same technique. In particular, we have

$$\Pr_{\mathcal{S}\leftarrow\mu^m}\left[\hat{\mu}_{\mathcal{S}}(\hat{\mathcal{E}}_\epsilon) \geq h(\mu, c, \eta, \alpha+\delta, \gamma+\delta', \epsilon, \mathcal{G}) + \delta\right] \leq 3\phi(m,\delta) + \phi_\epsilon(m,\delta). \tag{4.15}$$

Combining (4.14) and (4.15) by union bound completes the proof. □

Following Theorem 3.15, we can further show that if $\mathcal{G}$ also satisfies a universal approximation property, the optimal value of the empirical concentration problem (4.6) will approximately converge to the actual concentration function, if we increase both the complexity of the collection of subsets $\mathcal{G}$ and the number of samples used for the empirical estimation.

Theorem 4.7. *Consider the input metric probability space $(\mathcal{X}, \mu, \Delta)$, the concept function c and the label distribution function η. Let $\{\mathcal{G}(T)\}_{T\in\mathbb{N}}$ be a series of collection of subsets over $\mathcal{X}$. For any $T \in \mathbb{N}$, assume ϕ^T and ϕ^T_ϵ are complexity penalties for $\mathcal{G}(T)$ and $\mathcal{G}_\epsilon(T)$ respectively, and $\hat{\eta}$ is a function such that $\int_{\mathcal{X}} \|\hat{\eta}(\boldsymbol{x}) - \eta(\boldsymbol{x})\|_1 d\boldsymbol{x} \leq \delta_\eta$.*

Define $h(\mu, c, \eta, \alpha, \gamma, \epsilon, \mathcal{G}) = \inf_{\mathcal{E}\in\mathcal{G}}\{\mu(\mathcal{E}_\epsilon) : \mu(\mathcal{E}) \geq \alpha, \mathrm{LU}(\mathcal{E}; \mu, c, \eta) \geq \gamma\}$ as the constrained concentration function. We write $h(\mu, c, \eta, \alpha, \gamma, \epsilon)$ when $\mathcal{G} = \mathrm{Pow}(\mathcal{X})$. Given a sequence of datasets $\{S_T\}_{T\in\mathbb{N}}$, where S_T consists of $m(T)$ i.i.d. samples from μ and a sequence of numbers $\{\delta(T)\}_{T\in\mathbb{N}}$ with $\delta(T) \in (0, \alpha/2)$, if the following assumptions holds:

1. $\sum_{T=1}^{\infty} \phi^T(m(T), \delta(T)) < \infty$
2. $\sum_{T=1}^{\infty} \phi^T_\epsilon(m(T), \delta(T)) < \infty$

3. $\lim_{T\to\infty} \delta(T) = 0$

4. $\lim_{T\to\infty} h(\mu, c, \eta, \alpha, \gamma, \epsilon, \mathcal{G}(T)) = h(\mu, c, \eta, \alpha, \gamma, \epsilon)$[2]

5. *h is locally continuous w.r.t. α and γ at* $(\mu, c, \eta, \alpha, \gamma \pm \delta_\eta/\alpha, \epsilon, \text{Pow}(\mathcal{X}))$,

then with probability 1*, we have*

$$h(\mu, c, \eta, \alpha, \gamma - \delta_\eta/\alpha, \epsilon) \leq \lim_{T\to\infty} h(\mu_{\mathcal{S}_T}, c, \hat{\eta}, \alpha, \gamma, \epsilon, \mathcal{G}(T)) \leq h(\mu, c, \eta, \alpha, \gamma + \delta_\eta/\alpha, \epsilon).$$

Note that there is an error term of δ_η/α on the parameter γ. When the difference between the empirical label distribution $\hat{\eta}(\cdot)$ and the underlying label distribution $\eta(\cdot)$ is negligible, it is guaranteed that the optimal value of (4.6) asymptoptically converges to that of (4.5).

Proof of Theorem 4.7. Let E_T be the event such that

$$h\big(\mu, c, \eta, \alpha - \delta(T), \gamma - \delta'(T), \epsilon, \mathcal{G}(T)\big) - \delta(T) > h\big(\hat{\mu}_{\mathcal{S}_T}, c, \hat{\eta}, \alpha, \gamma, \epsilon, \mathcal{G}(T)\big) \text{ or}$$
$$h\big(\mu, c, \eta, \alpha + \delta(T), \gamma + \delta'(T), \epsilon, \mathcal{G}(T)\big) + \delta(T) < h\big(\hat{\mu}_{\mathcal{S}_T}, c, \hat{\eta}, \alpha, \gamma, \epsilon, \mathcal{G}(T)\big),$$

$\delta'(T) = (4\delta(T) + \delta_\eta)/(\alpha - 2\delta(T))$ for any $T \in \mathbb{N}$. Since $\delta(T) < \alpha/2$, thus according to Theorem 4.6, for any $T \in \mathbb{N}$, we have

$$\Pr[E_T] \leq 6\phi^T(m(T), \delta(T)) + 2\phi_\epsilon^T(m(T), \delta(T))).$$

By Assumptions 1 and 2, this further implies

$$\sum_{T=1}^{\infty} \Pr[E_T] \leq 6\sum_{T=1}^{\infty} \phi^T(m(T), \delta(T)) + 2\sum_{T=1}^{\infty} \phi_\epsilon^T(m(T), \delta(T))) < \infty.$$

[2]It is worth nothing that this assumption is satisfied for any family of collections of subsets that is a universal approximator, such as kernel SVMs and decision trees.

Thus according to Lemma 3.16, we know that there exists some $j \in \mathbb{N}$ such that for all $T \geq j$,

$$h\big(\mu, c, \eta, \alpha - \delta(T), \gamma - \delta'(T), \epsilon, \mathcal{G}(T)\big) - \delta(T) \leq h(\hat{\mu}_{\mathcal{S}_T}, c, \hat{\eta}, \alpha, \gamma, \epsilon) \leq h\big(\mu, c, \eta, \alpha + \delta(T), \gamma + \delta'(T), \epsilon, \mathcal{G}(T)\big) + \delta(T), \tag{4.16}$$

holds with probability 1. In addition, by Assumptions 3, 4 and 5, we have

$$\begin{aligned}
&\lim_{T\to\infty} h\big(\mu, c, \eta, \alpha - \delta(T), \gamma - \delta'(T), \epsilon, \mathcal{G}(T)\big) \\
&= \lim_{T_1\to\infty} \lim_{T_2\to\infty} h\big(\mu, c, \eta, \alpha - \delta(T_1), \gamma - \delta'(T_1), \epsilon, \mathcal{G}(T_2)\big) \\
&= \lim_{T_1\to\infty} h\big(\mu, c, \eta, \alpha - \delta(T_1), \gamma - \delta'(T_1), \epsilon\big) \\
&= h\big(\mu, c, \eta, \alpha, \gamma - \delta_\eta/\alpha, \epsilon\big),
\end{aligned}$$

where the second equality is due to Assumption 4 and the last equality is due to Assumptions 3 and 5. Similarly, we have

$$\lim_{T\to\infty} h\big(\mu, c, \eta, \alpha + \delta(T), \gamma + \delta'(T), \epsilon, \mathcal{G}(T)\big) = h\big(\mu, c, \eta, \alpha, \gamma + \delta_\eta/\alpha, \epsilon\big).$$

Therefore, let T goes to ∞ in (4.16), we have

$$h\big(\mu, c, \eta, \alpha, \gamma - \delta_\eta/\alpha, \epsilon\big) \leq \lim_{T\to\infty} h\big(\hat{\mu}_{\mathcal{S}_T}, c, \hat{\eta}, \alpha, \gamma, \epsilon, \mathcal{G}(T)\big) \leq h\big(\mu, c, \eta, \alpha, \gamma + \delta_\eta/\alpha, \epsilon\big),$$

which completes the proof. □

Concentration Estimation Algorithm. Although Theorem 4.7 provides a general idea how to choose $\mathcal{G}$ for measuring concentration, it does not indicate how to solve the empirical concentration problem (4.6) for a specific perturbation metric. This section presents a heuristic algorithm for estimating the least-expansive subset for optimization problem (4.6) when the metric is ℓ_2-norm or ℓ_∞-norm. We choose $\mathcal{G}$ as a union of balls for the ℓ_2-norm distance metric and set $\mathcal{G}$ as a union of hypercubes for ℓ_∞-norm. It is worth noting that such choices of $\mathcal{G}$ satisfy the condition required for Theorem 4.7, since they are universal set approximators and the VC-dimensions of both $\mathcal{G}$ and $\mathcal{G}_\epsilon$ are both bounded (see [36, 28]).

Algorithm 5: Heuristic Search for Robust Error Region under $\ell_p (p \in \{2, \infty\})$

Input : a set of labeled inputs $\{\boldsymbol{x}, c(\boldsymbol{x}), \hat{\eta}(\boldsymbol{x})\}_{\boldsymbol{x}\in\mathcal{S}}$, parameters $\alpha, \gamma, \epsilon, T$

$\hat{\mathcal{E}} \leftarrow \{\}, \quad \hat{\mathcal{S}}_{\text{init}} \leftarrow \{\}, \quad \hat{\mathcal{S}}_{\text{exp}} \leftarrow \{\}$;

for $t = 1, 2, \ldots, T$ **do**

  $k_{\text{lower}} \leftarrow \lceil(\alpha|\mathcal{S}| - |\hat{\mathcal{S}}_{\text{init}}|)/(T - t + 1)\rceil, \quad k_{\text{upper}} \leftarrow (\alpha|\mathcal{S}| - |\hat{\mathcal{S}}_{\text{init}}|)$;

  $\Omega \leftarrow \{\}$;

  for $\boldsymbol{u} \in \mathcal{S}$ **do**

    for $k \in [k_{lower}, k_{upper}]$ **do**

      $r_k(\boldsymbol{u}) \leftarrow$ compute the ℓ_p distance from $\boldsymbol{u}$ to the k-th nearest neighbour in $\mathcal{S} \setminus \hat{\mathcal{S}}_{\text{init}}$;

      $\mathcal{S}_{\text{init}}(\boldsymbol{u}, k) \leftarrow \{\boldsymbol{x} \in \mathcal{S} \setminus \hat{\mathcal{S}}_{\text{init}} : \|\boldsymbol{x} - \boldsymbol{u}\|_2 \le r_k(\boldsymbol{u})\}$;

      $\mathcal{S}_{\text{exp}}(\boldsymbol{u}, k) \leftarrow \{\boldsymbol{x} \in \mathcal{S} \setminus \hat{\mathcal{S}}_{\text{exp}} : \|\boldsymbol{x} - \boldsymbol{u}\|_2 \le r_k(\boldsymbol{u}) + \epsilon\}$;

      if $\text{LU}(\mathcal{S}_{init}(\boldsymbol{u}, k), \hat{\mu}_{\mathcal{S}}, c, \hat{\eta}) \ge \gamma$ **then**

        insert $(\boldsymbol{u}, k)$ into Ω

      end

    end

  end

  $(\hat{\boldsymbol{u}}, \hat{k}) \leftarrow \text{argmin}_{(\boldsymbol{u},k)\in\Omega}\{|\mathcal{S}_{\text{exp}}(\boldsymbol{u}, k)| - |\mathcal{S}_{\text{init}}(\boldsymbol{u}, k)|\}$;

  $\hat{\mathcal{E}} \leftarrow \hat{\mathcal{E}} \cup \text{Ball}(\hat{\boldsymbol{u}}, r_{\hat{k}}(\hat{\boldsymbol{u}}))$;

  $\hat{\mathcal{S}}_{\text{init}} \leftarrow \hat{\mathcal{S}}_{\text{init}} \cup \mathcal{S}_{\text{init}}(\hat{\boldsymbol{u}}, \hat{k}), \quad \hat{\mathcal{S}}_{\text{exp}} \leftarrow \hat{\mathcal{S}}_{\text{exp}} \cup \mathcal{S}_{\text{exp}}(\hat{\boldsymbol{u}}, \hat{k})$;

end

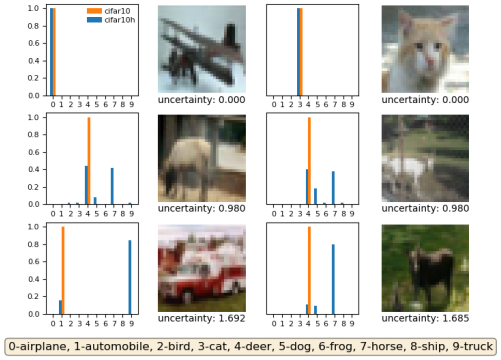
Output : $\hat{\mathcal{E}}$

The remaining task is to solve (4.6) based on the selected $\mathcal{G}$. We place the balls for ℓ_2 (or the

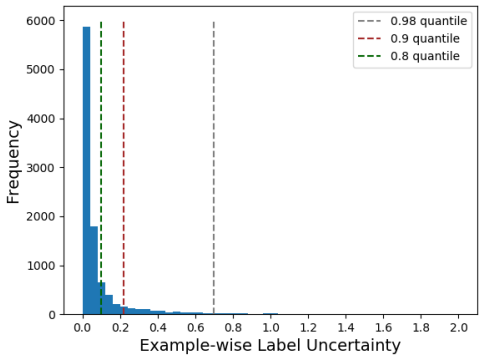
hypercubes for ℓ_∞) in a sequential manner, and search for the best placement that satisfies the label uncertainty constraint using a greedy approach. Algorithm 5 gives pseudocode for the search algorithm. It initializes the feasible set of the hyperparmeters Ω as an empty set for each placement of balls (or hypercubes), then enumerates all the possible initial placements, $\mathcal{S}_{\text{init}}(\boldsymbol{u}, k)$, such that its empirical label uncertainty exceeds the given threshold γ. Finally, among all the feasible ball (or hypercube) placements, it records the one that has the smallest ϵ-expansion with respect to the empirical measure $\hat{\mu}_{\mathcal{S}}$. In this way, the input region produced by Algorithm 5 serves as a good approximate solution to the empirical concentration problem (4.6).

4.5 Experiments

We conduct experiments on the CIFAR-10H dataset [95], which contains soft labels reflecting human perceptual uncertainty for the 10,000 CIFAR-10 test images [68]. 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 $c(\cdot)$. We report on experiments showing the connection between label uncertainty and classification error rates (Section 4.5.1) and that incorporating label uncertainty enables better intrinsic robustness estimates (Section 4.5.2). Section 4.5.3 demonstrates the possibility of improving model robustness by abstaining for inputs in high label uncertainty regions, whereas Section 4.5.4 explores whether a state-of-the-art confident learning approach can be used for predicting label uncertainty.



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



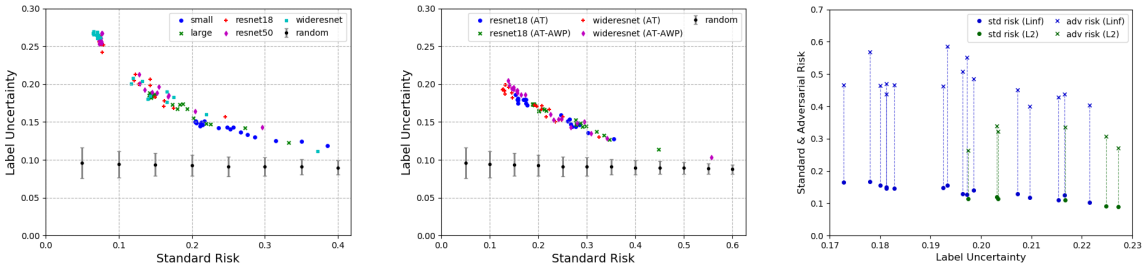
(b) Label Uncertainty Distribution

Figure 4.2: (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 4.3. (b) Histogram of the label uncertainty distribution for the CIFAR-10 test dataset.

4.5.1 Error Regions have Larger Label Uncertainty

Figure 4.2(a) shows the label uncertainty scores 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 (first row of Figure 4.2(a)), whereas images with high uncertainty scores look ambiguous or even misleading (second and third rows). Figure 4.2(b) shows the histogram of the label uncertainty distribution for all the $10,000$ CIFAR-10 test examples. In particular, more than 80% of the examples have label uncertainty scores below 0.1, suggesting the original class labels mostly capture the underlying label distribution well. However, around 2% of the examples have label uncertainty scores exceeding 0.7, and some 400 images appear to be mislabeled with uncertainty scores above 1.2.

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. That is, their induced error regions should have larger that typical label uncertainty. To test this hypothesis, we

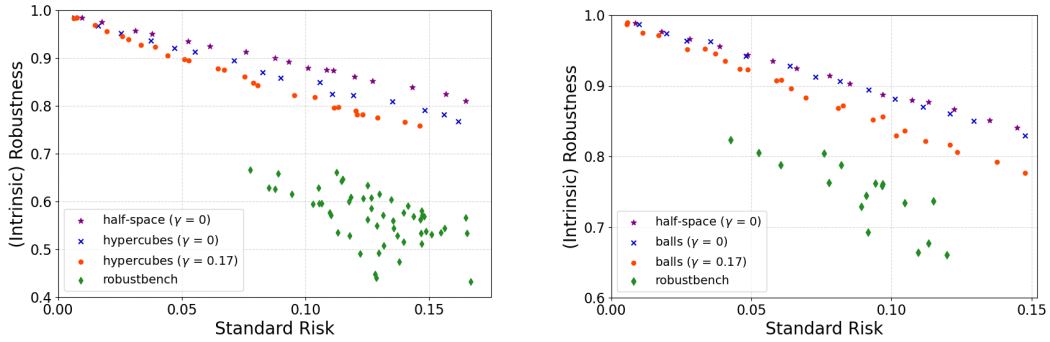


(a) Standard Training (b) Adversarial Training (c) RobustBench

Figure 4.3: 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.

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, adversarial risk, and label uncertainty of the corresponding error region.

Figures 4.3(a) and 4.3(b) demonstrate the relationship between label uncertainty and standard risk for various classifiers produced by standard training 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 4.3(c) summarizes the error region label uncertainty with respect to the state-of-the-art adversarially robust models documented in RobustBench [21]. 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



(a) ℓ_∞ perturbations ($\epsilon = 8/255$) (b) ℓ_2 perturbations ($\epsilon = 0.5$)

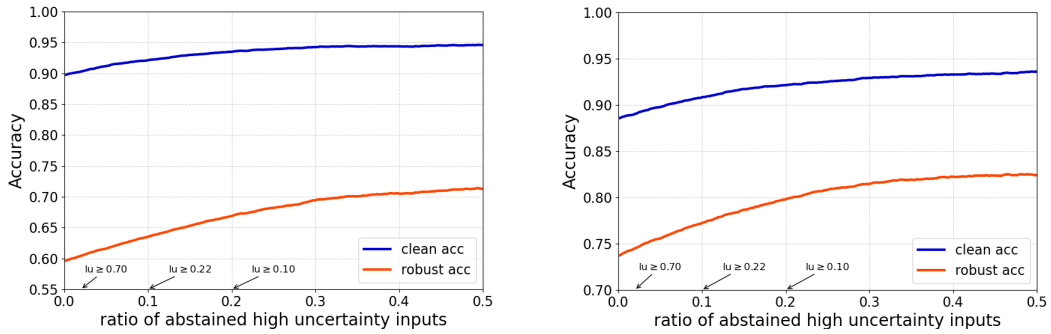
Figure 4.4: Estimated intrinsic robustness based on Algorithm 5 with $\gamma = 0.17$ under (a) ℓ_∞ perturbations with $\epsilon = 8/255$; and (b) ℓ_2 perturbations with $\epsilon = 0.5$. For comparison, we plot baseline estimates produced without considering label uncertainty using a half-space searching method [96] and using union of hypercubes or balls (Algorithm 5 with $\gamma = 0$). Robust accuracies achieved by state-of-the-art RobustBench models are plotted in green.

supports our hypothesis that error regions of state-of-the-art classifiers tend to have larger label uncertainty, and our claim that intrinsic robustness estimates should account for labels.

4.5.2 Empirical Estimation of Intrinsic Robustness

In this section, we apply Algorithm 5 to estimate the intrinsic robustness limit for the CIFAR-10 dataset under ℓ_∞ perturbations with $\epsilon = 8/255$ and ℓ_2 perturbations with $\epsilon = 0.5$. We set the label uncertainty threshold $\gamma = 0.17$ to roughly represent the error region label uncertainty of state-of-the-art classification models (see Figure 4.3). In particular, we adopt a $50/50$ train-test split over the original $10,000$ CIFAR-10 test images.

Figure 4.4 shows our intrinsic robustness estimates with $\gamma = 0.17$ when choosing different values of α. We include the estimates of intrinsic robustness defined with $\mathcal{F}_\alpha$ as a baseline, where no label uncertainty constraint is imposed ($\gamma = 0$). Results are shown both for our ℓ_p-balls searching method and the half-space searching method proposed in Section 3.5. We also plot the standard error and the robust accuracy of the state-of-the-art adversarially



(a) (b)

Figure 4.5: Accuracy curves for different adversarially-trained classifiers, varying the abstaining ratio of CIFAR-10 images with high label uncertainty score: (a) [17]'s model for ℓ_∞ perturbations with $\epsilon = 8/255$; (b) [123]'s model for ℓ_2 perturbations with $\epsilon = 0.5$. Corresponding cut-off values of label uncertainty are marked on the x-axis with respect to percentage values of $\{0.02, 0.1, 0.2\}$.

robust models in RobustBench [21]. For concentration estimation methods, the plotted values are the empirical measure of the returned optimally-searched subset (x-axis) and $1-$ the empirical measure of its ϵ-expansion (y-axis).

Compared with the baseline estimates, our label-uncertainty constrained intrinsic robustness estimates are uniformly lower across all the considered settings. Although both of these estimates can serve as legitimate upper bounds on the maximum achievable adversarial robustness for the given task, our estimate, which takes data labels into account, being closer to the robust accuracy achieved by state-of-the-art classifiers indicates it is a more accurate characterization of intrinsic robustness limit. For instance, under ℓ_∞ perturbations with $\epsilon = 8/255$, the best adversarially-trained classifier achieves 66% robust accuracy with approximately 8% clean error, whereas our estimate indicates that the maximum robustness one can hope for is about 82% as long as the model has at least 8% clean error. In contrast, the intrinsic robustness limit implied by standard concentration is as high as 90% for the same setting, which again shows the insufficiency of standard concentration.

4.5.3 Abstaining based on Label Uncertainty

According to our results presented in the previous sections, we expect classification models to have higher accuracy on examples with low label uncertainty. Figure 4.5 shows the results of experiments to study the effect of abstaining based on label uncertainty on both clean and robust accuracies using adversarially-trained CIFAR-10 classification models from [17] ($\ell_\infty, \epsilon = 8/255$) and [123] ($\ell_2, \epsilon = 0.5$). We first sort all the test CIFAR-10 images based on label uncertainty, then evaluate the model performance with respect to different abstaining ratios of top uncertain inputs. The accuracy curves suggest that a potential way to improve the robustness of classification systems is to enable the classifier an option to abstain on examples with high label uncertainty score.

For example, if we allow the robust classifier of [17] to abstain on the 2% of the test examples whose label uncertainty exceeds 0.7, the clean accuracy improves from 89.7% to 90.3%, while the robust accuracy increases from 59.5% to 60.4%. This is close to the maximum robust accuracy that could be achieved with a 2% abstention rate ($0.595/(1-0.02) = 0.607$). This result points to abstaining on examples in high label uncertainty regions as a promising path towards achieving adversarial robustness.

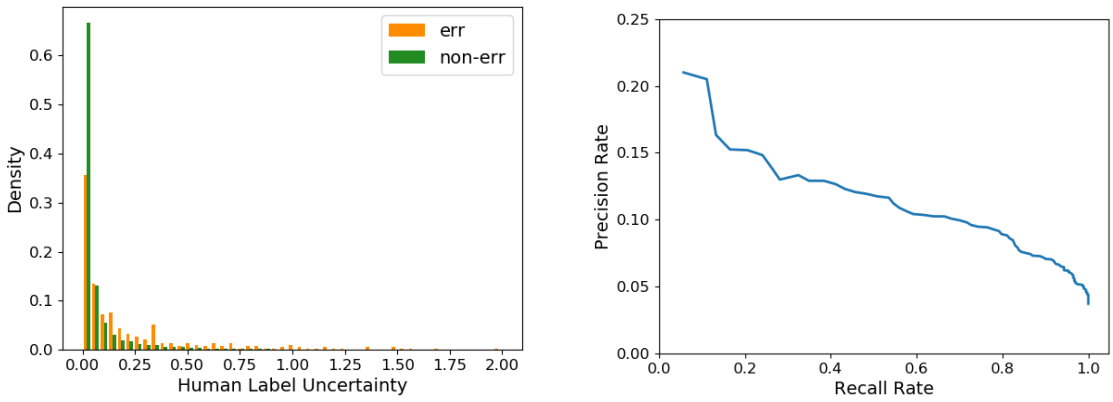
4.5.4 Estimating Label Errors using Confident Learning

The proposed concentration estimation framework relies on the knowledge of human soft labels to determine which example has label uncertainty exceeding a certainty threshold. Since typical datasets do not provide such label information like CIFAR-10H, this raises the question of how to extend our method to the setting where human soft labels are unavailable. In this section, we make an initial attempt to address the aforementioned issue using the confident learning approach of [91]. Their goal was to identify label errors for a

dataset, which is closely related to label uncertainty. The method first computes a confidence joint matrix based on the predicted probabilities of a pretrained classifier, then selects the top examples based on a ranking rule, such as self-confidence or max margin. If we are able to approximate human label uncertainty from the raw inputs and labels, or identify the set of examples with high label uncertainty, then we can immediately adapt our proposed framework by leveraging such estimated results. However, we observe only a weak correlation between the set of label errors that are produced by confident learning and the set of examples with high human label uncertainty.

We conduct the experiments on CIFAR-10 and identify the set of label errors based on confident learning. We train a ResNet-50 based classification model on the CIFAR-10 training data, and select examples in the CIFAR-10 test dataset as labeling errors using the best ranking method suggested in [91]. Figure 4.6(a) compares the distribution of human label uncertainty (based on the human soft labels from CIFAR-10H) between the set of estimated label error and non-errors. Although the set of examples estimated as label error have relative higher human label uncertainty compared with non-errors, there exist over 30% of estimated label errors have 0 label uncertainty for human annotators. It implies that there is a mismatch between label errors identified by human and that estimated using confident learning techniques. This is confirmed by the precision-recall curve presented in Figure 4.6(b). We treat examples with human label uncertainty exceeding 0.5 as the 'ground-truth' uncertain images, and vary the size of produced set of label errors to plot the precision and recall curve. The fact that precision rate is uniformly lower than 0.25, indicating that over 75% of the estimate error examples have human label uncertainty less than 0.5.

Figure 4.7 visualizes the human label distribution and estimated label distribution on CIFAR-10. We compute the estimated label uncertainty of each CIFAR-10 testing examples by replacing the human label distribution with the predicted probabilities of the trained model.



(a) Error versus Non-error (b) Precision Recall Curve

Figure 4.6: Illustration of misalignment label errors recognized by human and those identified by confident learning (a) Distribution of human label uncertainty between errors and non-errors estimated using confident learning; (b) Precision-recall curve for estimating the set of examples with human label uncertainty exceeding 0.5.

There exists a misalignment between the human label distribution and the distribution estimated using some neural network. This again confirms that label errors produced by confident learning are not guaranteed to be examples that are difficult for humans.

4.6 Summary

In this chapter, we show that standard concentration fails to sufficiently capture intrinsic robustness since it ignores data labels. Based on the definition of label uncertainty, we observe that the error regions induced by state-of-the-art classification models all tend to have high label uncertainty. This motivates us to develop an empirical method to study the concentration behavior regarding the input regions with high label uncertainty, which results in more accurate intrinsic robustness measures for benchmark image classification tasks. Our experiments show the importance of considering labels in understanding intrinsic robustness, and further suggest that abstaining based on label uncertainty could be a potential method to improve the classifier accuracy and robustness.

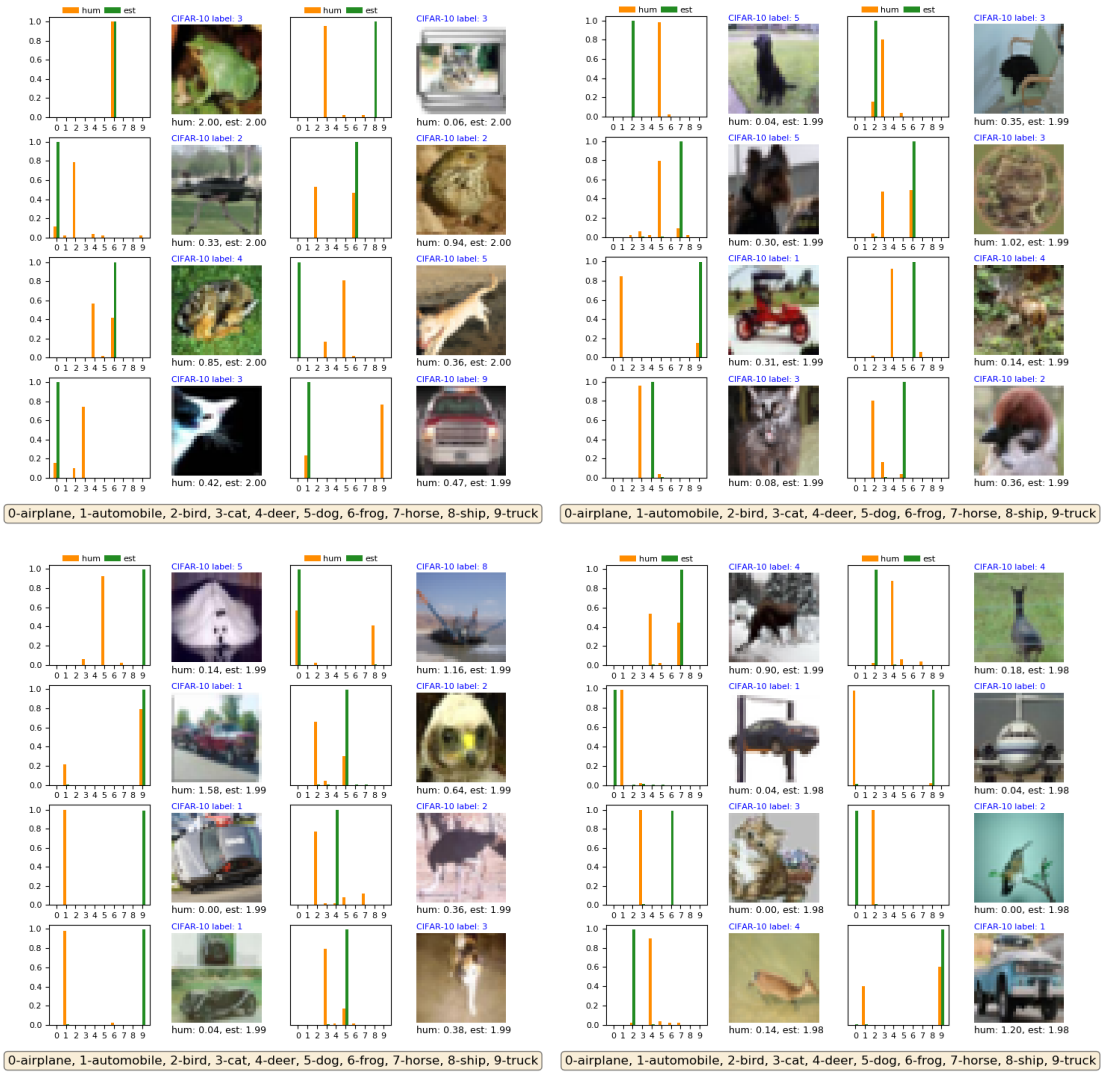


Figure 4.7: Visualization of label distribution of top uncertain CIFAR-10 test images estimated using a confident learning approach. Both human and estimated label distribution are plotted in each figure. The corresponding label uncertainty scores are computed and provided under each image, while the original CIFAR-10 label is highlighted in blue above each image.

Chapter 5

Learning Robust Representations[1]

5.1 Introduction

Motivated by the empirical and theoretical challenges of robust learning with adversarial examples, we study the underlying problem of learning adversarially robust representations [46, 94]. Given an input space $\mathcal{X} \subseteq \mathbb{R}^d$ and a feature space $\mathcal{Z} \subseteq \mathbb{R}^n$, any function $g : \mathcal{X} \to \mathcal{Z}$ is called a representation with respect to $(\mathcal{X}, \mathcal{Z})$. Adversarially robust representations denote the set of functions from $\mathcal{X}$ to $\mathcal{Z}$ that are less sensitive to adversarial perturbations with respect to some metric Δ. Note that one can always get an overall classification model by learning a downstream classifier given a representation, thus learning representations that are robust can be viewed as an intermediate step for the ultimate goal of finding adversarially robust models. In this sense, learning adversarially robust representations may help us better understand adversarial examples, and perhaps more importantly, bypass some of the aforementioned intrinsic barriers for achieving model robustness.

In the following, we first give a general definition for robust representations based on mutual information, then study its implications on model robustness for a downstream classification task. Finally, we propose empirical methods for estimating and inducing representation robustness, and demonstrate the effectiveness of our method through experiments.

[1] Sicheng Zhu*, Xiao Zhang*, David Evans, *Learning Adversarially Robust Representations via Worst-Case Mutual Information Maximization*, in the Thirty-seventh International Conference on Machine Learning (ICML 2022) [135].

5.2 Preliminaries

Mutual information. Mutual information is an entropy-based measure of the mutual dependence between variables.

Definition 5.1. Let (X, Z) be a pair of random variables with values over the space $\mathcal{X} \times \mathcal{Z}$. The *mutual information* of (X, Z) is defined as:

$$\mathrm{I}(X; Z) = \int_{\mathcal{Z}} \int_{\mathcal{X}} p_{XZ}(\boldsymbol{x}, \boldsymbol{z}) \log \left(\frac{p_{XZ}(\boldsymbol{x}, \boldsymbol{z})}{p_X(\boldsymbol{x}) p_Z(\boldsymbol{z})} \right) d\boldsymbol{x} d\boldsymbol{z},$$

where p_{XZ} is the joint probability density function of (X, Z), and p_X, p_Z are the marginal probability density functions of X and Z, respectively.

Intuitively, $\mathrm{I}(X; Z)$ tells us how well one can predict Z from X (and X from Z, since it is symmetrical). By definition, $\mathrm{I}(X; Z) = 0$ if X and Z are independent; when X and Z are identical, $\mathrm{I}(X; X)$ equals to the entropy $\mathrm{H}(X)$.

Wasserstein distance. Wasserstein distance is a distance function defined between two probability distributions on a given metric space.

Definition 5.2. Let $(\mathcal{X}, \Delta)$ be a metric space with bounded support. Given two probability measures μ and ν on $(\mathcal{X}, \Delta)$, the *p-th Wasserstein distance*, for any $p \geq 1$, is defined as:

$$\mathrm{W}_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} \Delta(\boldsymbol{x}, \boldsymbol{x}')^p \, d\gamma(x, x') \right)^{1/p},$$

where $\Gamma(\mu, \nu)$ is the collection of all probability measures on $\mathcal{X} \times \mathcal{X}$ with μ and ν being the marginals of the first and second factor, respectively. The *p-th Wasserstein ball* with respect

to μ and radius $\epsilon \geq 0$ is defined as:

$$\mathcal{B}_\epsilon(\mu; \mathrm{W}_p) = \{\mu' \in \mathcal{P}(\mathcal{X}) : \mathrm{W}_p(\mu', \mu) \leq \epsilon\}.$$

Note that the ∞-Wasserstein distance is defined as the limit of p-th Wasserstein distance, $\mathrm{W}_\infty(\mu, \nu) = \lim_{p\to\infty} \mathrm{W}_p(\mu, \nu)$.

Adversarial risk. Adversarial risk captures the vulnerability of a given classification model to input perturbations. In this chapter, we work with the following definition of adversarial risk, which has been used for robustness evaluation in previous works such as [79, 119, 126].

Definition 5.3. Let $(\mathcal{X}, \Delta)$ be the input metric space and $\mathcal{Y}$ be the set of labels. Let μ_{XY} be the underlying distribution of the input and label pairs. For any classifier $f : \mathcal{X} \to \mathcal{Y}$, the *adversarial risk* of f with respect to $\epsilon \geq 0$ is defined as:

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

Adversarial risk with $\epsilon = 0$ is equivalent to standard risk, namely $\mathrm{AdvRisk}_0(f) = \mathrm{Risk}(f) = \Pr_{(\boldsymbol{x},y)\sim\mu}[f(\boldsymbol{x}) \neq y]$. Note that different from the adversarial risk definition used in previous sections (Definition 3.1), Definition 5.3 does not assume the existence of a ground-truth concept function. However, as long as the small perturbations preserve the underlying ground-truth, these definitions of adversarial risk are equivalent to each other.

In addition, for any classifier $f : \mathcal{X} \to \mathcal{Y}$, we define the *adversarial gap* of f as:

$$\mathrm{AG}_\epsilon(f) = \mathrm{AdvRisk}_\epsilon(f) - \mathrm{Risk}(f).$$

5.3 Adversarially Robust Representations

In this section, we first propose a definition of representation vulnerability, and then prove a theorem that bounds achievable model robustness based on representation vulnerability.

5.3.1 Defining Representation Vulnerability

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space and $\mathcal{Z} \subseteq \mathbb{R}^n$ be some feature space. We define a representation to be a function g that maps any input $\boldsymbol{x}$ in $\mathcal{X}$ to some vector $g(\boldsymbol{x}) \in \mathcal{Z}$. A classifier, $f = h \circ g$, maps an input to a label in a label space $\mathcal{Y}$, and is a composition of a downstream classifier, $h : \mathcal{Z} \to \mathcal{Y}$, with a representation, $g : \mathcal{X} \to \mathcal{Z}$. As is done in previous works [46, 61], we define a feature as a function from $\mathcal{X}$ to $\mathbb{R}$, so can think of a representation as an array of features.

Inspired by the empirical success of standard representation learning using the mutual information maximization principle [58], we propose the following definition of *representation vulnerability*, which captures the robustness of a given representation against input distribution perturbations in terms of mutual information between its input and output.

Definition 5.4 (Representation Vulnerability)**.** Let $(\mathcal{X}, \mu_X, \Delta)$ be a metric probability space of inputs and $\mathcal{Z}$ be some feature space. Given a representation $g : \mathcal{X} \to \mathcal{Z}$ and $\epsilon \geq 0$, the *representation vulnerability* of g with respect to perturbations bounded in an ∞-Wasserstein ball with radius ϵ is defined as:

$$\mathrm{RV}_\epsilon(g) = \sup_{\mu_{X'} \in \mathcal{B}_\epsilon(\mu_X; \mathrm{W}_\infty)} \big[\mathrm{I}(X; g(X)) - \mathrm{I}(X'; g(X'))\big],$$

where X and X' denote random variables that follow μ_X and $\mu_{X'}$, respectively.

Representation vulnerability is always non-negative, and higher values indicate that the representation is less robust to adversarial input distribution perturbations. More formally, given parameters $\epsilon \geq 0$ and $\tau \geq 0$, a representation g is called (ϵ, τ)-*robust* if $\text{RV}_\epsilon(g) \leq \tau$.

Notably, using the ∞-Wasserstein distance does not restrict the choice of the metric function Δ of the input space. This metric Δ corresponds to the perturbation metric for defining adversarial examples. Thus, based on our definition of representation vulnerability, our following theoretical results and empirical methods work with any adversarial perturbation, including any ℓ_p-norm based attack.

Compared with existing definitions of robust features [46, 61, 43], our definition is more general and enjoys several desirable properties. As it does not impose any constraint on the feature space, it is invariant to scale change[2] and it does not require the knowledge of the labels. The most similar definition to ours is from [94], who propose to use statistical Fisher information as the evaluation criteria for feature robustness. However, Fisher information can only capture the average sensitivity of the log conditional density to small changes on the input distribution (when $\epsilon \to 0$), whereas our definition is defined with respect to the worst-case input distribution perturbations in an ∞-Wasserstein ball, which is more aligned with the adversarial setting. As will be shown next, our representation robustness notion has a clear connection with the potential model robustness of any classifier that can be built upon a representation.

5.3.2 Theoretical Results

In this section, we present our main theoretical results regarding robust representations. First, we present the following lemma that characterizes the connection between adversarial

[2]Scale-invariance is desirable for representation robustness. Otherwise, one can always divide the function by some large constant to improve its robustness, e.g., [46].

risk and input distribution perturbations bounded in an ∞-Wasserstein ball.

Lemma 5.5. *Let $(\mathcal{X}, \Delta)$ be the input metric space and $\mathcal{Y}$ be the set of labels. Assume all the examples are generated from a joint probability distribution $(X, Y) \sim \mu_{XY}$. Let μ_X be the marginal distribution of X. Then, for any classifier $f : \mathcal{X} \to \mathcal{Y}$ and $\epsilon > 0$, we have*

$$\text{AdvRisk}_\epsilon(f) = \sup_{\mu_{X'} \in \mathcal{B}_\epsilon(\mu_X; \mathrm{W}_\infty)} \Pr\left[f(X') \neq Y\right],$$

where X' denotes the random variable that follows $\mu_{X'}$.

Proof of Lemma 5.5. Before starting the proof, we introduce the following notations and an alternative definition of the ∞-Wasserstein distance. Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces. We say that $T : \mathcal{X} \to \mathcal{Y}$ transports $\mu \in \mathcal{P}(\mathcal{X})$ to $\nu \in \mathcal{P}(\mathcal{Y})$, and we call T a transport map, if $\nu(\mathcal{A}) = \mu(T^{-1}(\mathcal{A}))$, for all ν-measurable sets $\mathcal{E}$. In addition, for any measurable map $T : \mathcal{X} \to \mathcal{Y}$, we define the *pushforward* of μ through T as

$$(T_\#(\mu))(\mathcal{A}) = \mu(T^{-1}(\mathcal{A})), \quad \text{for any measurable } \mathcal{A} \subseteq \mathcal{Y}.$$

Alternative definition of ∞-Wasserstein distance. From the perspective of transportation theory, given two probability measures μ and ν on $(\mathcal{X}, \Delta)$, any joint probability distribution $\gamma \in \Gamma(\mu, \nu)$ corresponds to a specific transport map $T : \mathcal{X} \to \mathcal{X}$ that moves μ to ν. Then, the p-th Wasserstein distance can viewed as finding the optimal transport map to move from μ to ν that minimizes some cost functional depending on p [67]. For the case where $p = \infty$, if we let T be the transport map induced by a given $\gamma \in \Gamma(\mu, \nu)$, then the cost functional can be informally understood as the maximum of all the transport distances $\Delta(T(\boldsymbol{x}), \boldsymbol{x})$.

More rigorously, the ∞-Wasserstein distance can be alternatively defined as

$$\begin{aligned}\mathrm{W}_\infty(\mu,\nu) &:= \inf_{\gamma\in\Gamma(\mu,\nu)} \gamma\text{-ess sup}_{(\boldsymbol{x},\boldsymbol{x}')\in\mathcal{X}^2} \Delta(\boldsymbol{x},\boldsymbol{x}') \\ &= \inf_{\gamma\in\Gamma(\mu,\nu)} \inf\big\{t\geq 0\colon \gamma\big(\Delta(\boldsymbol{x},\boldsymbol{x}')>t\big)=0\big\}.\end{aligned}$$

A more detailed discussion of ∞-Wasserstein distance can be found in [18].

Now we are ready to prove Lemma 5.5. We are going prove the equality by proving $\leq$ inequalities in both directions. First, we prove

$$\mathrm{AdvRisk}_\epsilon(f) \leq \sup_{\mu_{X'}\in\mathcal{B}_{\mathrm{W}_\infty}(\mu_X,\epsilon)} \Pr\big[f(X')\neq Y\big]. \tag{5.1}$$

For any classifier $f:\mathcal{X}\to\mathcal{Y}$, according to Definition 3.1, we have

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

Since f is a given deterministic function, the optimal perturbation scheme that achieves $\mathrm{AdvRisk}_\epsilon(f)$ essentially defines a transport map $T:\mathcal{X}\to\mathcal{X}$. More specifically, let $\mathcal{C}_y(f)=\{\boldsymbol{x}\in\mathcal{X}: f(\boldsymbol{x})\neq y\}$. Then, for any sampled pair $(\boldsymbol{x},y)\sim\mu_{XY}$, we can construct T such that

$$T(\boldsymbol{x}) = \begin{cases} \mathrm{argmin}_{\boldsymbol{x}'\in\mathcal{C}_y(f)}\Delta(\boldsymbol{x}',\boldsymbol{x}), & \text{if } \mathcal{C}_y(f)\cap\mathcal{B}(\boldsymbol{x},\epsilon)\neq\varnothing; \\ \boldsymbol{x}, & \text{otherwise.}\end{cases}$$

Let (X,Y) be the random variable that follows μ_{XY}. By construction, it can be easily verified that $T_\#(\mu_X)\in\mathcal{B}_\epsilon(\mu_X;\mathrm{W}_\infty)$ and $\mathrm{AdvRisk}_\epsilon(f)=\Pr\big[f(T(X))\neq Y\big]$. Therefore, we have proven (5.1).

It remains to prove the other direction of the inequality:

$$\text{AdvRisk}_\epsilon(f) \geq \sup_{\mu_{X'} \in \mathcal{B}_{\text{w}_\infty}(\mu_X, \epsilon)} \Pr\left[f(X') \neq Y\right]. \tag{5.2}$$

According to the alternative definition of ∞-Wasserstein distance, the optimal solution $\mu^*_{X'}$ that achieves the supremum of the right hand side of (5.2) can be captured by a transport map $T^* : \mathcal{X} \to \mathcal{X}$ such that $\mu^*_{X'} = T^*_\#(\mu_X)$ and $\Delta(T^*(X), X) \leq \epsilon$ holds almost surely with respect to the randomness of X and T^*. Thus, we have

$$\begin{aligned}
\Pr\left[f(T^*(X)) \neq Y\right] &= \Pr_{(\boldsymbol{x},y)\sim\mu_{XY}}\left[f(T^*(\boldsymbol{x})) \neq y\right] \\
&= \Pr_{(\boldsymbol{x},y)\sim\mu_{XY}}\left[\Delta(T^*(\boldsymbol{x}), \boldsymbol{x}) \leq \epsilon \text{ and } f(T^*(\boldsymbol{x})) \neq y\right] \\
&\leq 1 - \Pr_{(\boldsymbol{x},y)\sim\mu_{XY}}\left[\forall\, \boldsymbol{x}' \in \mathcal{B}_\epsilon(\boldsymbol{x}) \text{ s.t. } f(\boldsymbol{x}') = y\right] = \text{AdvRisk}_\epsilon(f).
\end{aligned}$$

Therefore, we have proven the second direction and completed the proof. □

The following theorem gives a lower bound for the adversarial risk for any downstream classifier, using the worst-case mutual information between the representation's input and output distributions.

Theorem 5.6. *Let $(\mathcal{X}, \Delta)$ be the input metric space, $\mathcal{Y}$ be the set of labels and μ_{XY} be the underlying joint probability distribution. Assume the marginal distribution of labels μ_Y is a uniform distribution over $\mathcal{Y}$. Consider the feature space $\mathcal{Z}$ and the set of downstream classifiers $\mathcal{H} = \{h : \mathcal{Z} \to \mathcal{Y}\}$. Given $\epsilon \geq 0$, for any $g : \mathcal{X} \to \mathcal{Z}$, we have*

$$\inf_{h\in\mathcal{H}} \text{AdvRisk}_\epsilon(h \circ g) \geq 1 - \frac{\text{I}(X;Z) - \text{RV}_\epsilon(g) + \log 2}{\log|\mathcal{Y}|},$$

where X is the random variable that follows the marginal distribution of inputs μ_X and

$Z = g(X)$.

Proof of Theorem 5.6. Before starting the proof, we state two useful lemmas on Markov chains. A Markov chain is defined to be a collection of random variables $\{X_t\}_{t\in\mathbb{Z}}$ with the property that given the present, the future is conditionally independent of the past. Namely,

$$\Pr(X_t = j | X_0 = i_0, X_1 = i_1, ..., X_{(t-1)} = i_{(t-1)}) = \Pr(X_t = j | X_{(t-1)} = i_{(t-1)}).$$

Lemma 5.7 (Fano's Inequality)**.** *Let X be a random variable uniformly distributed over a finite set of outcomes $\mathcal{X}$. For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$ forms a Markov chain, we have*

$$\Pr(\hat{X} \neq X) \geq 1 - \frac{\mathrm{I}(X;\hat{X}) - \log 2}{\log|\mathcal{X}|}.$$

Lemma 5.8 (Data-Processing Inequality)**.** *For any Markov chain $X \to Y \to Z$, we have*

$$\mathrm{I}(X;Y) \geq I(X;Z) \quad \textit{and} \quad \mathrm{I}(Y;Z) \geq \mathrm{I}(X;Z).$$

Chapter 2 in [20] provides proofs of Lemmas 5.7 and 5.8.

Now we are ready to prove Theorem 5.6.

For any classifier $h : \mathcal{Z} \to \mathcal{Y}$, according to Lemma 5.5, we have

$$\mathrm{AdvRisk}_\epsilon(h \circ g) = \sup_{\mu_{X'} \in \mathcal{B}_{\mathrm{W}_\infty}(\mu_X, \epsilon)} \Pr\left[h(g(X')) \neq Y\right]. \tag{5.3}$$

Let $\mu_{X'} \in \mathcal{B}_\epsilon(\mu_X; \mathrm{W}_\infty)$ be a probability measure over $(\mathcal{X}, \Delta)$. According to the alternative definition of ∞-Wasserstein distance using optimal transport, $\mu_{X'}$ corresponds to a transport

map $T : \mathcal{X} \to \mathcal{X}$ such that $\mu_{X'} = T_{\#}(\mu_X)$. Thus, for any given $\mu_{X'} \in \mathcal{B}_\epsilon(\mu_X; \mathrm{W}_\infty)$ and $h \in \mathcal{H}$, we have the Markov chain

$$Y \to X \xrightarrow{T} X' \xrightarrow{g} g(X') \xrightarrow{h} (h \circ g)(X').$$

where X, Y are random variables for input and label distributions respectively. The first Markov chain $Y \to X$ can be understood as a generative model for generating inputs according to the conditional probability distribution $\mu_{X|Y}$. Therefore, applying Lemmas 5.7 and 5.8, we obtain the inequality,

$$\Pr\big[h(g(X')) \neq Y\big] \geq 1 - \frac{\mathrm{I}\big(Y; (h \circ g)(X')\big) + \log 2}{\log |\mathcal{Y}|} \geq 1 - \frac{\mathrm{I}\big(X'; g(X')\big) + \log 2}{\log |\mathcal{Y}|}. \quad (5.4)$$

Taking the supremum over the distribution of X' in $\mathcal{B}_\epsilon(\mu_X; \mathrm{W}_\infty)$ and infimum over $h \in \mathcal{H}$ on both sides of (5.4) yields

$$\begin{aligned}
\inf_{h \in \mathcal{H}} \big[\mathrm{AdvRisk}_\epsilon(h \circ g)\big] &= \inf_{h \in \mathcal{H}} \sup_{\mu_{X'} \in \mathcal{B}_\epsilon(\mu_X; \mathrm{W}_\infty)} \Pr\big[h(g(X')) \neq Y\big] \\
&\geq 1 - \frac{\inf_{\mu_{X'} \in \mathcal{B}_\epsilon(\mu_X; \mathrm{W}_\infty)} \mathrm{I}\big(X'; g(X')\big) + \log 2}{\log |\mathcal{Y}|} \\
&= 1 - \frac{\mathrm{I}(X; g(X)) - \mathrm{RV}_\epsilon(g) + \log 2}{\log |\mathcal{Y}|},
\end{aligned}$$

where the first equality is due to (5.3) and the inequality holds because of (5.4). Thus, we completed the proof. □

Remark 5.9. Theorem 5.6 suggests that adversarial robustness cannot be achieved if the available representation is highly vulnerable or the standard mutual information between X and $g(X)$ is low. Note that $\mathrm{I}(X; g(X)) - \mathrm{RV}_\epsilon(g) = \inf\{\mathrm{I}(X'; g(X')) : X' \sim \mu_{X'} \in \mathcal{B}_\epsilon(\mu_X; \mathrm{W}_\infty)\}$, which corresponds to the worst-case mutual information between input and output of g. Therefore, if we assume robust classification as the downstream task for repre-

sentation learning, then the representation having high worst-case mutual information is a necessary condition for achieving adversarial robustness for the overall classifier.

In addition, it is worth noting that Theorem 5.6 can be extended to general p-th Wasserstein distances, if the downstream classifiers are evaluated based on robustness under distributional shift[3], instead of adversarial risk. To be more specific, if using W_p metric to define representation vulnerability, we can then establish an upper bound on the maximum distributional robustness with respect to the considered W_p metric for any downstream classifier based on similar proof techniques of Theorem 5.6.

5.4 Measuring Representation Vulnerability

This section presents an empirical method for estimating the vulnerability of a given representation using i.i.d. samples. Recall from Definition 5.4, for any $g : \mathcal{X} \to \mathcal{Z}$, the representation vulnerability of g with respect to the input metric probability space $(\mathcal{X}, \mu_X, \Delta)$ and $\epsilon \geq 0$ is defined as:

$$\mathrm{RV}_\epsilon(g) = \underbrace{\mathrm{I}(X; g(X))}_{J_1} - \underbrace{\inf_{\mu_{X'} \in \mathcal{B}_\epsilon(\mu_X; \mathrm{W}_\infty)} \mathrm{I}(X'; g(X'))}_{J_2} . \tag{5.5}$$

To measure representation vulnerability, we need to compute both terms J_1 and J_2. However, the main challenge is that we do not have the knowledge of the underlying probability distribution μ_X for real-world problem tasks. Instead, we only have access to a finite set of data points sampled from the distribution. Therefore, it is natural to consider sample-based estimator for J_1 and J_2 for practical use.

The first term J_1 is essentially the mutual information between X and $Z = g(X)$. A variety

[3]See [113] for a rigorous definition of distributional robustness.

of methods have been proposed for estimating mutual information [87, 24, 115, 63, 86]. The most effective estimator is the mutual information neural estimator (MINE) [8], based on the dual representation of KL-divergence [33]:

$$\hat{\mathrm{I}}_m(X;Z) = \sup_{\theta\in\Theta} \mathbb{E}_{\hat{\mu}_{XZ}^{(m)}}[T_\theta] - \log\left(\mathbb{E}_{\hat{\mu}_X^{(m)}\otimes\hat{\mu}_Z^{(m)}}[\exp(T_\theta)]\right),$$

where $T_\theta : \mathcal{X}\times\mathcal{Z} \to \mathbb{R}$ is the function parameterized by a deep neural network with parameters $\theta \in \Theta$, and $\hat{\mu}_{XZ}^{(m)}$, $\hat{\mu}_X^{(m)}$ and $\hat{\mu}_Z^{(m)}$ denote the empirical distributions[4] of random variables (X, Z), X and Z respectively, based on m samples. In addition, [8] empirically demonstrates the superiority of the proposed estimator in terms of estimation accuracy and efficiency, and prove that it is strongly consistent: for all $\varepsilon > 0$, there exists $M \in \mathbb{Z}$ such that for any $m \geq M$, $|\hat{\mathrm{I}}_m(X;Z) - \mathrm{I}(X;Z)| \leq \varepsilon$ almost surely. Given the established effectiveness of this method, we implement MINE to estimate $\mathrm{I}(X; g(X))$ as the first step.

Compared with J_1, the second term J_2 is much more difficult to estimate, as it involves finding the worst-case perturbations on μ_X in a ∞-Wasserstein ball in terms of mutual information. As with the estimation of J_1, we only have a finite set of instances sampled from μ_X. On the other hand, due to the non-linearity and the lack of duality theory with respect to the ∞-Wasserstein distance [18], it is inherently difficult to directly solve an ∞-Wasserstein constrained optimization problem, even if we work with the empirical distribution of μ_X. To deal with the first challenge, we replace μ_X with its empirical measure $\hat{\mu}_X^{(m)}$ based on i.i.d. samples. Then, to avoid the need to search through the whole ∞-Wasserstein ball, we restrict the search space of $\mu_{X'}$ to be the following set of empirical distributions:

$$\mathcal{A}_\epsilon(\mathcal{S}) = \left\{\frac{1}{m}\sum_{i=1}^m \delta_{\boldsymbol{x}_i'} : \boldsymbol{x}_i' \in \mathcal{B}_\epsilon(\boldsymbol{x}_i)\ \forall i \in [m]\right\}, \tag{5.6}$$

[4]Given a set of m samples $\{\boldsymbol{x}_i\}_{i\in[m]}$ from a distribution μ, we let $\hat{\mu}^{(m)} = \frac{1}{m}\sum_{i\in[m]}\delta_{\boldsymbol{x}_i}$ be the empirical measure of μ.

where $\mathcal{S} = \{\boldsymbol{x}_i : i \in [m]\}$ denotes the given set of m data points sampled from μ_X. Note that the considered set $\mathcal{A}_\epsilon(\mathcal{S}) \subseteq \mathcal{B}_\epsilon(\hat{\mu}_X^{(m)}; \mathrm{W}_\infty)$, since each perturbed point $\boldsymbol{x}_i'$ is at most ϵ-away from $\boldsymbol{x}_i$. Finally, making use of the dual formulation of KL-divergence that is used in MINE, we propose the following empirical optimization problem for estimating J_2:

$$\min_{\mu_{X'}} \hat{\mathrm{I}}_m\big(X'; g(X')\big) \;\text{ s.t. }\; \mu_{X'} \in \mathcal{A}_\epsilon(\mathcal{S}), \tag{5.7}$$

where we simply set the empirical distribution $\hat{\mu}_{X'}^{(m)}$ to be the same as $\mu_{X'}$. In addition, we propose a heuristic alternating minimization algorithm to solve (5.7) (see Appendix B of [135] for the pseudocode and a complexity analysis of the proposed algorithm). More specifically, our algorithm alternatively performs gradient ascent on θ for the inner maximization problem of estimating $\hat{\mathrm{I}}_m(X'; g(X'))$ given μ_X', and searches for the set of worst-case perturbations on $\{\boldsymbol{x}_i' : i \in [m]\}$ given θ based on projected gradient descent.

5.5 Learning Robust Representations

In this section, we present our method for learning adversarially robust representations. First, we introduce the mutual information maximization principle for representation learning [75, 9]. Mathematically, given an input probability distribution μ_X and a set of representations $\mathcal{G} = \{g : \mathcal{X} \to \mathcal{Z}\}$, the maximization principle proposes to solve this problem:

$$\max_{g \in \mathcal{G}} \; \mathrm{I}\big(X; g(X)\big). \tag{5.8}$$

Although this principle has been shown to be successful for learning good representations under the standard setting [58], it becomes ineffective when considering adversarial perturbations (see Table 5.1 for an illustration). Motivated by the theoretical connections be-

tween feature sensitivity and adversarial risk for downstream robust classification shown in Section 5.3, we stimulate robust representations by adding a regularization term based on representation vulnerability:

$$\max_{g \in \mathcal{G}} \quad \mathrm{I}(X; g(X)) - \beta \cdot \mathrm{RV}_{\epsilon}(g), \tag{5.9}$$

where $\beta \geq 0$ is the trade-off parameter between $\mathrm{I}(X; g(X))$ and $\mathrm{RV}_{\epsilon}(g)$. When $\beta = 0$, (5.9) is same as the objective for learning standard representations (5.8). Increasing the value of β will produce representations with lower vulnerability, but may undesirably affect the standard mutual information $\mathrm{I}(g(X); X)$ if β is too large. In particular, we set $\beta = 1$ in the following discussions, which allows us to simplify (5.9) to obtain the following problem:

$$\max_{g \in \mathcal{G}} \min_{\mu_{X'} \in \mathcal{B}_{\epsilon}(\mu_X; \mathrm{W}_{\infty})} \mathrm{I}\big(X'; g(X')\big). \tag{5.10}$$

The proposed training principle (5.10) aims to maximize the mutual information between the representation's input and output under the worst-case input distribution perturbation bounded in a ∞-Wasserstein ball. We remark that problem (5.10) aligns well with the results of Theorem 5.6, which shows the importance of the learned feature representation achieving high worst-case mutual information for a downstream robust classification task.

As with estimating the feature sensitivity in Section 5.4, we do not have access to the underlying μ_X. However, the inner minimization problem is exactly the same as estimating the worst-case mutual information J_2 in (5.5), thus we can simply adapt the proposed empirical estimator (5.7) to solve (5.10). To be more specific, we reparameterize g using a neural network with parameter $\psi \in \Psi$ and use the following min-max optimization problem:

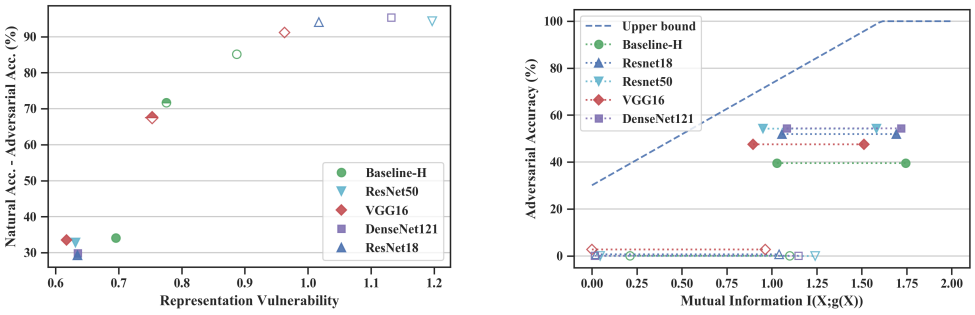
$$\max_{\psi \in \Psi} \min_{\mu_{X'} \in \mathcal{A}_{\epsilon}(\mathcal{S})} \hat{\mathrm{I}}_m\big(X'; g_{\psi}(X')\big). \tag{5.11}$$

Based on the proposed algorithm for the inner minimization problem, (5.11) can be efficiently solved using a standard optimizer, such as stochastic gradient descent.

5.6 Experiments

This section reports on experiments to study the implications of robust representations on benchmark image datasets. Instead of focusing directly improving model robustness, our experiments focus on understanding the proposed definition of robust representations as well as its implications. Based on the proposed estimator in Section 5.4, Section 5.6.1 summarizes experiments to empirically test the relationship between representation vulnerability and model robustness, by extracting internal representations from the state-of-the-art pre-trained standard and robust classification models. In addition, we empirically evaluate the general lower bound on adversarial risk presented in Theorem 5.6. In Section 5.6.2, we evaluate the proposed training principle for learning robust representations on image datasets, and test its performance with comparisons to the state-of-the-art standard representation learning method in a downstream robust classification framework. We also visualize saliency maps as an intuitive criteria for evaluating representation robustness.

We conduct experiments on CIFAR-10 [68], considering typical ℓ_∞-norm bounded adversarial perturbations with $\epsilon = 8/255$. We use the PGD attack [79] for both generating adversarial distributions in the estimation of worst-case mutual information and evaluating model robustness. To implement our proposed estimator (5.7), we adopt the *encode-and-dot-product* model architecture in [58] and adjust it to adapt to different forms of representations. We leverage implementations from [38] and [58] in our implementation.



(a) (b)

Figure 5.1: (a) Normal and worst case mutual information for logit-layer representations. Each pair of points shows the result of a specific model—the left point indicates the worst case mutual information and the right for the normal mutual information. Filled points are robust models; hollow points are standard models. (b) Correlations between the representation vulnerability and the CIFAR-10 model's natural-adversarial accuracy gap. Filled points indicate robust models (trained with $\epsilon = 8/255$), half-filled are models adversarially trained with $\epsilon = 2/255$, and unfilled points are standard models.

5.6.1 Representation Robustness

To evaluate our proposed definition on representation vulnerability and its implications for downstream classification models, we conduct experiments on image benchmarks using various classifiers, including VGG [111], ResNet [55], DenseNet [59] and the simple convolutional neural network in [58] denoted as Baseline-H.

Correlation with model robustness. We empirically evaluate the correlation between our representation vulnerability definition and achievable model robustness on image benchmarks. Figure 5.1(a) summarizes the results of these experiments for CIFAR-10, where we set the logit layer as the considered representation space. The adversarial gap decreases with decreasing representation vulnerability in an approximately consistent relationship. Models with low logit layer representation vulnerability tend to have low natural-adversarial accuracy gap, which is consistent with the intuition behind our definition.

Adversarial risk lower bound. Theorem 5.6 provides a lower bound on the adversarial risk that can be achieved by any downstream classifier as a function of representation vulnerability. To evaluate the tightness of this bound, we estimate the normal-case and worst-case mutual information $\mathrm{I}(X; g(X))$ of layer representation g for different models, and empirically evaluate the adversarial risk of the models. Figure 5.1(b) shows the results, where we again set the logit layer as the feature space for a more direct comparison. The lower bound of adversarial risk is calculated according to Theorem 5.6 and is converted to the upper bound of adversarial accuracy for reference. In particular, for standard models, both the estimated worst-case mutual information and the adversarial accuracy are close to zero, whereas the computed upper bounds on adversarial accuracy are around 30%. We empirically observed around 50% adversarial accuracy for robust models, whereas the bounds computed using the estimated worst-case mutual information and Theorem 5.6 are about 75%. This shows that Theorem 5.6 gives a reasonably tight bound for a model's adversarial accuracy with respect to the logit-layer representation robustness.

Figure 5.1(b) also indicates that even the robust models produced by adversarial training have representations that are not sufficiently robust to enable robust downstream classifications. For example, robust DenseNet121 in our evaluations has the highest logit layer worst-case mutual information of 1.08, yet the corresponding adversarial accuracy is upper bounded by 77.0% which is unsatisfactory for CIFAR-10. Such information theoretic limitation also justifies our training principle of worst-case mutual information maximization, since on the other hand the adversarial accuracy upper bound calculated by normal-case mutual information does not constitute a limitation for most robust models in our experiments

(as in Figure 5.1(a), most robust models achieve adversarial accuracy close to 100%).

Internal feature robustness. We further investigate the implications of our proposed definition from the level of individual features. Specifically for neural networks, we consider the function from the input to each individual neuron within a layer as a feature. The motivations for considering feature robustness comes from the fact that mutual information in terms of the whole representation is controlled by the sum of all the features' mutual information and robust features are potentially easier to train [46]. As an illustration, we evaluate the robustness of all the convolutional kernels in the second layer of the Baseline-H model. Each neuron evaluated here is a composite convolutional kernel (all kernels in the first layer connected to a second layer kernel) with image input size 10×10. Figure 5.2 shows the results that are averaged over two independently trained models for each type. This result reveals the apparent difference in feature robustness between a standard model and the adversarially-trained robust model, even in lower layers. Although in this case the result does not prohibit a robust downstream model for lower layers neurons, for neurons in higher layers the difference becomes more distinct and the vulnerability of neurons can thus be the bottleneck of achieving high model robustness. The different feature robustness according to our definition also coincide with the saliency maps of features, where the saliency maps of robust features are apparently more interpretable compared to those of standard features.

5.6.2 Learning Robust Representations

Our worst-case mutual information maximization training principle provides an unsupervised way to learn adversarially robust representations. Since there are no established ways to measure the robustness of a representation, empirically testing the robustness of repre-

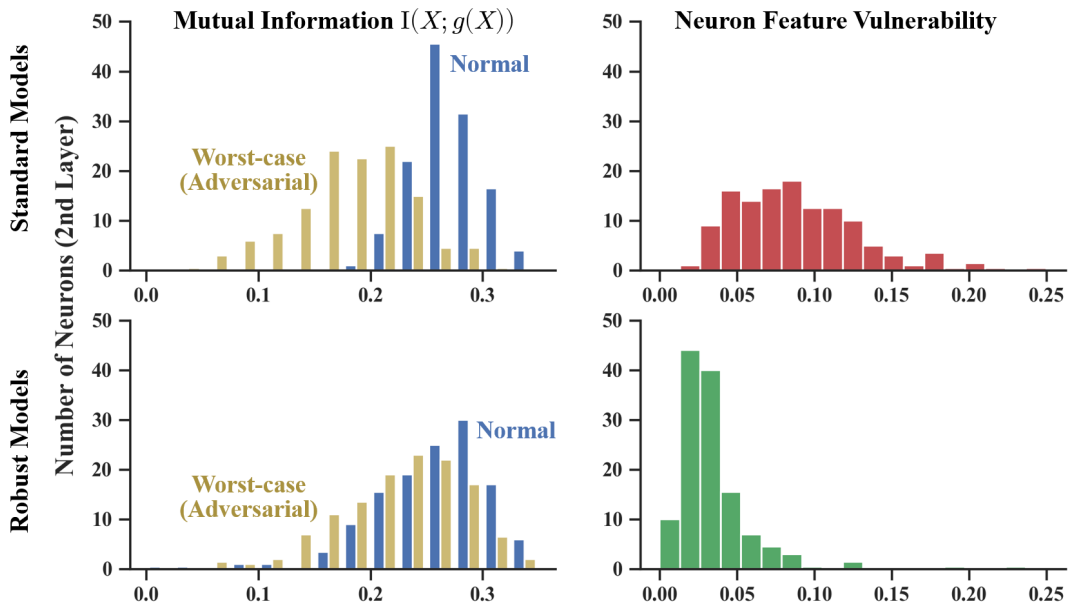


Figure 5.2: Distribution of mutual information $I(X; g(X))$ and feature vulnerability in the second convolutional layer of Baseline-H. The upper plots are for standard models, and the lower plots are for robust models. The total number of neurons is 128.

sentations learned by our training principle poses a dilemma. To avoid circular reasoning, we evaluate the learned representations by running a series of downstream adversarial classification tasks and comparing the performance of the best models we are able to find for each representation. In addition, recent work shows that the interpretability of saliency map has certain connections with robustness [41, 61], thus we study the saliency map as an alternative criteria for evaluating robust representations.

The unsupervised representation learning approach based on mutual information maximization principle in [58] achieves the state-of-the-art results in many downstream tasks, including standard classification. We further adopt their encoder architecture in our implementation, and extend their evaluation settings to adversarially robust classification. Specifically, we truncate the front part of Baseline-H with a 64-dimensional latent layer output as the representation g and train it by the worst-case mutual information maximization principle using only unlabeled data (removing the labels from the normal training data). We test two

		MLP h		Linear h	
g	h	Natural	Adversarial	Natural	Adversarial
[58]	Std.	58.77 ± 0.22	0.22 ± 0.08	47.01 ± 0.53	0.15 ± 0.03
[58]	Rob.	29.75 ± 1.49	15.08 ± 0.63	22.79 ± 1.42	10.28 ± 0.52
Ours	Std.	$\mathbf{62.54 \pm 0.12}$	14.06 ± 0.69	$\mathbf{50.29 \pm 0.58}$	10.98 ± 0.49
Ours	Std. (E.S.)	51.59 ± 3.34	27.53 ± 0.81	48.55 ± 0.63	13.52 ± 0.16
Ours	Rob.	52.34 ± 0.17	$\mathbf{31.52 \pm 0.31}$	43.55 ± 0.10	$\mathbf{25.15 \pm 0.10}$
Supervised Std.		86.33 ± 0.17	0.07 ± 0.02	86.36 ± 0.13	0.02 ± 0.01
Supervised Rob.		70.71 ± 0.58	40.50 ± 0.27	72.44 ± 0.59	39.98 ± 0.16

Table 5.1: Comparisons of different methods on CIFAR-10 in downstream classification settings. *E.S.* denotes early stopping under the criterion of the best adversarial accuracy. We present mean accuracy and the standard deviation over 4 repeated trials.

architectures (two-layer multilayer perceptron and linear classifier) for implementing the downstream classifier h and train it using labeled data after the encoder g has been trained using unlabeled data.

Downstream classification tasks. Comparison results on CIFAR-10 are demonstrated in Table 5.1. The fully-supervised models are trained for reference, from which we can see the simple model architecture we use achieves a decent natural accuracy of 86.3%; the adversarially-trained robust model reduces accuracy to around 70% with adversarial accuracy of 40.5%. The baseline, with g and h both trained normally, resembles the setting in [58] and achieves a natural accuracy of 58.8%. For representations learned using worst-case mutual information maximization, the composition with standard two-layer multilayer perceptron (MLP) h achieves a non-trivial (compared to the 0.2% for the standard representation) adversarial accuracy of 14.1%. When h is further trained using adversarial training, the robust accuracy increases to 31.5% which is comparable to the result of the robust fully-supervised model. As an ablation, the robust h based on standard g achieves an adversarial accuracy of 15.1%, yet the natural accuracy severely drops below 30%, indicating that a robust classifier cannot be found using the vulnerable representation. The

case where h is a simple linear classifier shows similar results. These comparisons show that the representation learned using worst-case mutual information maximization can make the downstream classification more robust over the baseline and approaches the robustness of fully-supervised adversarial training. This provides evidence that our training principle produces adversarially robust representations.

Another interesting implication given by results in Table 5.1 is that robustly learned representations may also have better natural accuracy (62.5%) over the standard representation (58.8%) in downstream classification tasks on CIFAR-10. This matches our experiments in Figure 5.1(b) where logit layer representations in robust models conveys more normal-case mutual information (up to 1.75) than those in standard models (up to 1.25).

Saliency maps. A saliency map is commonly defined as the gradient of a model's loss with respect to the model's input [41]. For a classification model, it intuitively illustrates what the model looks for in changing its classification decision for a given sample. Recent work [41, 61] indicates, at least in some synthetic settings, that the more alignment the saliency map has with the input image, the more adversarially robust the model is. As an additional test of representation robustness, we calculate the saliency maps of standard and robust representations g by the mutual information maximization loss with respect to the input. Figure 5.3 shows that the saliency maps of the robust representation appear to be much less noisy and more interpretable in terms of the alignment with original images. Intuitively, this shows that robust representations capture relatively higher level visual concepts instead of pixel-level statistical clues [39]. The more interpretable saliency maps of representation learned by our training principle further support its effectiveness in learning adversarially robust representation.

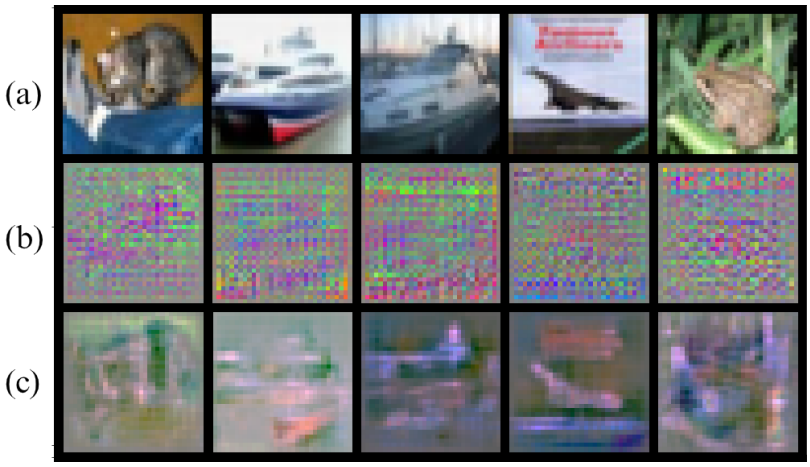


Figure 5.3: Visualization of saliency maps of different models on CIFAR-10: (a) original images (b) representations learned using [58] (c) representations learned using our method.

5.7 Summary

In this chapter, we proposed a novel definition of representation robustness based on the worst-case mutual information, and showed both theoretical and empirical connections between our definition of representation robustness and model robustness for a downstream classification task. In addition, by developing estimation and training methods for representation robustness, we demonstrated the connection and the usefulness of the proposed method on benchmark datasets. Our results are not enough to produce strongly robust models, but they provide a new approach for understanding and measuring achievable adversarial robustness at the level of representations.

Chapter 6

Towards Building Better Robust Models

6.1 Introduction

Previous chapters mainly focus on advancing our understanding on adversarial examples as well as studying the fundamental causes behind the adversarial vulnerability of existing machine learning classifiers. In this chapter, we shift our focus from understanding adversarial robustness to exploring ways to build better robust systems by rethinking the robustness design goal.

We show in Chapter 3 and Chapter 4 that there exists intrinsic limits for achieving adversarial robustness, due to the concentration of data distribution and the existence of inputs with high label uncertainty. Built upon the current design goal for adversarial robustness, we have no hope to escape such intrinsic robustness limits, unless we can produce a perfect classifier with zero standard risk. Learning such a perfect classifier, however, seems to be a difficult task, since uncertain inputs are likely to incur classification errors, and there may exist inherent trade-offs between robustness and accuracy [119, 99].

We argue that the typical design goal for building adversarially robust classifiers, which is termed as *overall robustness* in this chapter, may not be appropriate under certain scenarios. More formally, the overall robustness of a classifier f with respect to perturbations with

strength ϵ measured by metric Δ is defined as:

$$\text{AdvRob}_{\epsilon}(f) = 1 - \Pr_{\boldsymbol{x}\sim\mu}\left[\exists\, \boldsymbol{x}' \in \mathcal{B}_{\epsilon}(\boldsymbol{x}) \text{ s.t. } f(\boldsymbol{x}') \neq c(\boldsymbol{x})\right], \tag{6.1}$$

where μ represents the underlying input distribution and c is the concept function that assign a label to each input. For practical evaluation of overall robustness, the empirical measure based on a set of inputs $\{\boldsymbol{x}_i\}_{i\in[N]}$ is typically used to replace the population measure μ. Note that this notion of overall robustness is different from Definition 3.1, which is used throughout our previous discussions on intrinsic robustness. However, since we usually do not have the knowledge of the ground-truth label beyond standard inputs, it is much easier to measure the system's robustness based on overall robustness from a practical perspective.

Although overall robustness seems to reflect a classifier's resilience to adversarial perturbations, we argue in this chapter that it is less meaningful to be treated as the evaluation criterion for robustness when only certain kinds of adversarial misclassifications provide value for potential adversaries (Section 6.2), or when input examples with uncertain class labels are being assessed (Section 6.3). Therefore, we design more meaningful ways for assessing system's robustness performance and discuss potential methods for building better robust models under each of the aforementioned scenarios.

6.2 Cost-Sensitive Robustness[1]

6.2.1 Background

This section provides a brief introduction on related topics, including neural network classifiers, adversarial examples, defenses with certified robustness, and cost-sensitive learning.

Neural Network Classifiers. A K-layer neural network classifier can be represented by a function $f : \mathcal{X} \to \mathcal{Y}$ such that $f(\boldsymbol{x}) = f_{K-1}(f_{K-2}(\cdots(f_1(\boldsymbol{x}))))$, for any $\boldsymbol{x} \in \mathcal{X}$. For $k \in \{1, 2, \ldots, K-2\}$, the mapping function $f_k(\cdot)$ typically consists of two operations: an affine transformation (either matrix multiplication or convolution) and a nonlinear activation. In this paper, we consider rectified linear unit (ReLU) as the activation function. If denote the feature vector of the k-th layer as $\boldsymbol{z}_k$, then $f_k(\cdot)$ is defined as:

$$\boldsymbol{z}_{k+1} = f_k(\boldsymbol{z}_k) = \max\{\mathbf{W}_k \boldsymbol{z}_k + \boldsymbol{b}_k, \mathbf{0}\}, \quad \forall k \in \{1, 2, \ldots K-2\},$$

where $\mathbf{W}_k$ denotes the weight parameter matrix and $\boldsymbol{b}_k$ the bias vector. The output function $f_{K-1}(\cdot)$ maps the feature vector in the last hidden layer to the output space $\mathcal{Y}$ solely through matrix multiplication: $\boldsymbol{z}_K = f_{K-1}(\boldsymbol{z}_{K-1}) = \mathbf{W}_{K-1}\boldsymbol{z}_{K-1} + \boldsymbol{b}_{K-1}$, where $\boldsymbol{z}_K$ can be regarded as the estimated score vector of input $\boldsymbol{x}$ for different possible output classes. In the following discussions, we use f_θ to represent the neural network classifier, where $\theta = \{\mathbf{W}_1, \ldots, \mathbf{W}_{K-1}, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_{K-1}\}$ denotes the model parameters.

To train the neural network, a loss function $\sum_{i=1}^{N} \mathcal{L}(f_\theta(\boldsymbol{x}_i), y_i)$ is defined for a set of training examples $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ is the i-th input vector and y_i denotes its class label. Cross-entropy loss is typically used for multiclass image classification. With proper initialization,

[1] Xiao Zhang, David Evans, *Cost-Sensitive Robustness against Adversarial Examples*, in the Seventh International Conference on Learning Representations (ICLR 2019) [132].

all model parameters are then updated iteratively using backpropagation. For any input example $\tilde{\boldsymbol{x}}$, the predicted label $\hat{y}$ is given by the index of the largest predicted score among all classes, $\text{argmax}_j[f_\theta(\tilde{\boldsymbol{x}})]_j$.

Adversarial Examples. An adversarial example is an input, generated by some adversary, which is visually indistinguishable from an example from the natural distribution, but is able to mislead the target classifier. Since "visually indistinguishable" depends on human perception, which is hard to define rigorously, we consider the most popular alternative: input examples with perturbations bounded in ℓ_∞-norm [51]. The set of adversarial examples with respect to seed example $\{\boldsymbol{x}_0, y_0\}$ and classifier $f_\theta(\cdot)$ is defined as:

$$\mathcal{A}_\epsilon(\boldsymbol{x}_0, y_0; \theta) = \big\{\boldsymbol{x} \in \mathcal{X} : \|\boldsymbol{x} - \boldsymbol{x}_0\|_\infty \leq \epsilon \text{ and } \operatorname*{argmax}_j [f_\theta(\boldsymbol{x})]_j \neq y_0\big\}, \tag{6.2}$$

where $\epsilon > 0$ denotes the maximum perturbation distance. Although ℓ_p distances are commonly used in adversarial examples research, they are not an adequate measure of perceptual similarity [109] and other minimal geometric transformations can be used to find adversarial examples [40, 62, 124]. Nevertheless, there is considerable interest in improving robustness in this simple domain, and hope that as this research area matures we will find ways to apply results from studying simplified problems to more realistic ones.

Defenses with Certified Robustness. A line of recent work has proposed defenses that are guaranteed to be robust against norm-bounded adversarial perturbations. [56] proved formal robustness guarantees against ℓ_2-norm bounded perturbations for two-layer neural networks, and provided a training method based on a surrogate robust bound. [97] developed an approach based on semidefinite relaxation for training certified robust classifiers, but was limited to two-layer fully-connected networks. Our work builds most directly on [121], which can be applied to deep ReLU-based networks and achieves the state-of-the-art

certified robustness on MNIST dataset.

Following the definitions in [121], an adversarial polytope $\mathcal{Z}_\epsilon(\boldsymbol{x})$ with respect to a given example $\boldsymbol{x}$ is defined as

$$\mathcal{Z}_\epsilon(\boldsymbol{x}) = \big\{ f_\theta(\boldsymbol{x}+\boldsymbol{\delta}) : \|\boldsymbol{\delta}\|_\infty \leq \epsilon \big\}, \tag{6.3}$$

which contains all the possible output vectors for the given classifier f_θ by perturbing $\boldsymbol{x}$ within an ℓ_∞-norm ball with radius ϵ. A seed example, $\{\boldsymbol{x}_0, y_0\}$, is said to be *certified robust* with respect to maximum perturbation distance ϵ, if the corresponding adversarial example set $\mathcal{A}_\epsilon(\boldsymbol{x}_0, y_0; \theta)$ is empty. Equivalently, if we solve, for any output class $y_{\text{targ}} \neq y_0$, the optimization problem,

$$\underset{\boldsymbol{z}_K}{\text{minimize}} \;\; [\boldsymbol{z}_K]_{y_0} - [\boldsymbol{z}_K]_{y_{\text{targ}}}, \quad \text{subject to} \;\; \boldsymbol{z}_K \in \mathcal{Z}_\epsilon(\boldsymbol{x}_0), \tag{6.4}$$

then according to the definition of $\mathcal{A}_\epsilon(\boldsymbol{x}_0, y_0; \theta)$ in (6.2), $\{\boldsymbol{x}_0, y_0\}$ is guaranteed to be robust provided that the optimal objective value of (6.4) is positive for every output class. To train a robust model on a given dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$, the standard robust optimization aims to minimize the sample loss function on the worst-case locations through the following adversarial loss

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{N} \max_{\|\boldsymbol{\delta}\|_\infty \leq \epsilon} \mathcal{L}\big(f_\theta(\boldsymbol{x}_i+\boldsymbol{\delta}), y_i\big), \tag{6.5}$$

where $\mathcal{L}(\cdot,\cdot)$ denotes the cross-entropy loss. However, due to the nonconvexity of the neural network classifier $f_\theta(\cdot)$ introduced by the nonlinear ReLU activation, both the adversarial polytope (6.3) and training objective (6.5) are highly nonconvex. In addition, solving optimization problem (6.4) for each pair of input example and output class is computationally

intractable.

Instead of solving the optimization problem directly, [121] proposed an alternative training objective function based on convex relaxation, which can be efficiently optimized through a dual network. Specifically, they relaxed $\mathcal{Z}_\epsilon(\boldsymbol{x})$ into a convex outer adversarial polytope $\tilde{\mathcal{Z}}_\epsilon(\boldsymbol{x})$ by replacing the ReLU inequalities for each neuron $z = \max\{\hat{z}, 0\}$ with a set of inequalities,

$$z \geq 0, \quad z \geq \hat{z}, \quad -u\hat{z} + (u - \ell)z \leq -u\ell, \tag{6.6}$$

where u, ℓ denote the lower and upper bounds on the considered pre-ReLU activation.[2] Based on the relaxed outer bound $\tilde{\mathcal{Z}}_\epsilon(\boldsymbol{x})$, they propose the following alternative optimization problem,

$$\underset{\boldsymbol{z}_K}{\text{minimize}} \ [\boldsymbol{z}_K]_{y_0} - [\boldsymbol{z}_K]_{y_{\text{targ}}}, \quad \text{subject to } \boldsymbol{z}_K \in \tilde{\mathcal{Z}}_\epsilon(\boldsymbol{x}_0), \tag{6.7}$$

which is in fact a linear program. Since $\mathcal{Z}_\epsilon(\boldsymbol{x}) \subseteq \tilde{\mathcal{Z}}_\epsilon(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{X}$, solving (6.7) for all output classes provides stronger robustness guarantees compared with (6.4), provided all the optimal objective values are positive. In addition, they derived a guaranteed lower bound, denoted by $J_\epsilon\big(\boldsymbol{x}_0, g_\theta(\boldsymbol{e}_{y_0} - \boldsymbol{e}_{y_{\text{targ}}})\big)$, on the optimal objective value of Equation 6.7 using duality theory, where $g_\theta(\cdot)$ is a K-layer feedforward dual network (Theorem 1 in [121]). Finally, according to the properties of cross-entropy loss, they minimize the following objective to train the robust model, which serves as an upper bound of the adversarial loss (6.5):

$$\underset{\theta}{\text{minimize}} \ \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\bigg(- J_\epsilon\big(\boldsymbol{x}_i, g_\theta(\boldsymbol{e}_{y_i} \cdot \mathbf{1}^\top - \mathbf{I})\big), y_i\bigg), \tag{6.8}$$

[2]The elementwise activation bounds can be computed efficiently using Algorithm 1 in [121].

where $\mathbf{1}$ denotes the all-ones vector, $\mathbf{I}$ denotes the identity matrix and $g_\theta(\cdot)$ is regarded as a column-wise function when applied to a matrix. Although the proposed method in [121] achieves certified robustness, its computational complexity is quadratic with the network size in the worst case so it only scales to small networks. Recently, [122] extended the training procedure to scale to larger networks by using nonlinear random projections. However, if the network size allows for both methods, we observe a small decrease in performance using the training method provided in [122]. Therefore, we only use the approximation techniques for the experiments on CIFAR-10 (§6.2.3), and use the less scalable method for the MNIST experiments (§6.2.3).

Cost-Sensitive Learning. Cost-sensitive learning [32, 37, 77] was proposed to deal with unequal misclassification costs and class imbalance problems commonly found in classification applications. The key observation is that cost-blind learning algorithms tend to overwhelm the major class, but the neglected minor class is often our primary interest. For example, in medical diagnosis misclassifying a rare cancerous lesion as benign is extremely costly. Various cost-sensitive learning algorithms [70, 128, 134, 64] have been proposed in literature, but only a few algorithms, limited to simple classifiers, considered adversarial settings.[3] [23] studied the naive Bayes classifier for spam detection in the presence of a cost-sensitive adversary, and developed an adversary-aware classifier based on game theory. [4] proposed a cost-sensitive robust minimax approach that hardens a linear discriminant classifier with robustness in the adversarial context. All of these methods are designed for simple linear classifiers, and cannot be directly extended to neural network classifiers. In addition, the robustness of their proposed classifier is only examined experimentally based on the performance against some specific adversary, so does not provide any notion of certified robustness. Recently, [34] advocated for the idea of using application-level semantics

[3]Given the vulnerability of standard classifiers to adversarial examples, it is not surprising that standard cost-sensitive classifiers are also ineffective against adversaries.

in adversarial analysis, however, they didn't provide a formal method on how to train such classifier. Our work provides a practical training method that hardens neural network classifiers with certified cost-sensitive robustness against adversarial perturbations.

6.2.2 Training a Cost-Sensitive Robust Classifier

The approach introduced in [121] penalizes all adversarial class transformations equally, even though the consequences of adversarial examples usually depends on the specific class transformations. Here, we provide a formal definition of cost-sensitive robustness and propose a general method for training cost-sensitive robust models.

Certified Cost-Sensitive Robustness. Our approach uses a cost matrix $\mathbf{C}$ that encodes the cost (i.e., potential harm to model deployer) of different adversarial examples. First, we consider the case where there are m classes and $\mathbf{C}$ is a $m \times m$ binary matrix with $C_{jj'} \in \{0, 1\}$. The value $C_{jj'}$ indicates whether we care about an adversary transforming a seed input in class j into one recognized by the model as being in class j'. If the adversarial transformation $j \to j'$ matters, $C_{jj'} = 1$, otherwise $C_{jj'} = 0$. Let $\Omega_j = \{j' \in [m] : C_{jj'} \neq 0\}$ be the index set of output classes that induce cost with respect to input class j. For any $j \in [m]$, let $\delta_j = 0$ if Ω_j is an empty set, and $\delta_j = 1$ otherwise. We are only concerned with adversarial transformations from a seed class j to target classes $j' \in \Omega_j$. For any example $\boldsymbol{x}$ in seed class j, $\boldsymbol{x}$ is said to be *certified cost-sensitive robust* if the lower bound $J_\epsilon(\boldsymbol{x}, g_\theta(\boldsymbol{e}_j - \boldsymbol{e}_{j'})) \geq 0$ for all $j' \in \Omega_j$. That is, no adversarial perturbations in an ℓ_∞-norm ball around $\boldsymbol{x}$ with radius ϵ can mislead the classifier to any target class in Ω_j.

The *cost-sensitive robust error* on a dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$ is defined as the number of examples that are not guaranteed to be cost-sensitive robust over the number of non-zero cost candidate

seed examples:

$$\textit{cost-sensitive robust error} = 1 - \frac{\#\{i \in [N] : J_\epsilon(\boldsymbol{x}_i, g_\theta(\boldsymbol{e}_{y_i} - \boldsymbol{e}_{j'})) \geq 0, \forall j' \in \Omega_{y_i}\}}{\sum_{j|_{\delta_j=1}} N_j},$$

where $\#\mathcal{A}$ represents the cardinality of a set $\mathcal{A}$, and N_j is the total number of examples in class j.

Next, we consider a more general case where $\mathbf{C}$ is a $m \times m$ real-valued cost matrix. Each entry of $\mathbf{C}$ is a non-negative real number, which represents the cost of the corresponding adversarial transformation. To take into account the different potential costs among adversarial examples, we measure the cost-sensitive robustness by the average certified cost of adversarial examples. The cost of an adversarial example $\boldsymbol{x}$ in class j is defined as the sum of all $C_{jj'}$ such that $J_\epsilon(\boldsymbol{x}, g_\theta(\boldsymbol{e}_j - \boldsymbol{e}_{j'})) < 0$. Intuitively speaking, an adversarial example will induce more cost if it can be adversarially misclassified as more target classes with high cost. Accordingly, the *robust cost* is defined as the total cost of adversarial examples divided by the total number of valued seed examples:

$$\textit{robust cost} = \frac{\sum_{j|_{\delta_j=1}} \sum_{i|_{y_i=j}} \sum_{j' \in \Omega_j} C_{jj'} \cdot \mathbb{1}\left(J_\epsilon(\boldsymbol{x}_i, g_\theta(\boldsymbol{e}_j - \boldsymbol{e}_{j'})) < 0\right)}{\sum_{j|_{\delta_j=1}} N_j}, \tag{6.9}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function.

Cost-Sensitive Robust Optimization. Recall that our goal is to develop a classifier with certified cost-sensitive robustness, while maintaining overall classification accuracy. According to the guaranteed lower bound, $J_\epsilon\big(\boldsymbol{x}_0, g_\theta(\boldsymbol{e}_{y_0} - \boldsymbol{e}_{y_{\text{targ}}})\big)$ on Equation 6.7 and inspired by the cost-sensitive CE loss [64], we propose the following robust optimization with respect

to a neural network classifier f_θ:

$$\underset{\theta}{\text{minimize}}\ \frac{1}{N}\sum_{i\in[N]}\mathcal{L}\big(f_\theta(\boldsymbol{x}_i), y_i\big) + \alpha\sum_{j\in[m]}\frac{\delta_j}{N_j}\sum_{i|_{y_i=j}}\log\left(1+\sum_{j'\in\Omega_j}C_{jj'}\cdot\exp\big(-J_\epsilon(\boldsymbol{x}_i, g_\theta(\boldsymbol{e}_j-\boldsymbol{e}_{j'}))\big)\right), \tag{6.10}$$

where $\alpha \geq 0$ denotes the regularization parameter. The first term in Equation 6.10 denotes the cross-entropy loss for standard classification, whereas the second term accounts for the cost-sensitive robustness. Compared with the overall robustness training objective function (6.8), we include a regularization parameter α to control the trade-off between classification accuracy on original inputs and adversarial robustness.

To provide cost-sensitivity, the loss function selectively penalizes the adversarial examples based on their cost. For binary cost matrixes, the regularization term penalizes every cost-sensitive adversarial example equally, but has no impact for instances where $C_{jj'} = 0$. For the real-valued costs, a larger value of $C_{jj'}$ increases the weight of the corresponding adversarial transformation in the training objective. This optimization problem (6.10) can be solved efficiently using gradient-based algorithms, such as stochastic gradient descent and ADAM [66].

6.2.3 Experiments

We evaluate the performance of our cost-sensitive robustness training method on models for two benchmark image classification datasets: MNIST [73] and CIFAR-10 [68]. We compare our results for various cost scenarios with overall robustness training as a baseline. For both datasets, the relevant family of attacks is specified as all the adversarial perturbations

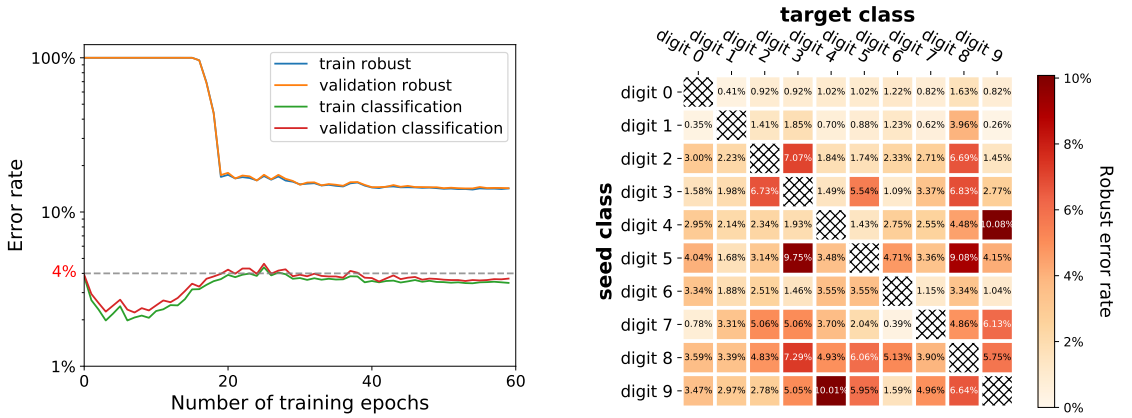
that are bounded in an ℓ_∞-norm ball.

Our goal in the experiments is to evaluate how well a variety of different types of cost matrices can be supported. MNIST and CIFAR-10 are toy datasets, thus there are no obvious cost matrices that correspond to meaningful security applications for these datasets. Instead, we select representative tasks and design cost matrices to capture them.

Experiments on MNIST.

For MNIST, we use the same convolutional neural network architecture [72] as [121], which includes two convolutional layers, with 16 and 32 filters respectively, and a two fully-connected layers, consisting of 100 and 10 hidden units respectively. ReLU activations are applied to each layer except the last one. For both our cost-sensitive robust model and the overall robust model, we randomly split the 60,000 training samples into five folds of equal size, and train the classifier over 60 epochs on four of them using the Adam optimizer [66] with batch size 50 and learning rate 0.001. We treat the remaining fold as a validation dataset for model selection. In addition, we use the ϵ-scheduling and learning rate decay techniques, where we increase ϵ from 0.05 to the desired value linearly over the first 20 epochs and decay the learning rate by 0.5 every 10 epochs for the remaining epochs.

Baseline: Overall Robustness. Figure 6.1(a) illustrates the learning curves of both classification error and overall robust error during training based on robust loss (6.8) with maximum perturbation distance $\epsilon = 0.2$. The model with classification error less than 4% and minimum overall robust error on the validation dataset is selected over the 60 training epochs. The best classifier reaches 3.39% classification error and 13.80% overall robust error on the 10,000 MNIST testing samples. We report the robust test error for every adversarial transformation in Figure 6.1(b) (for the model without any robustness training all of the values are 100%). The (i, j)-th entry is a bound on the robustness of that seed-



(a) learning curves

(b) heatmap of robust test error

Figure 6.1: Preliminary results on MNIST using overall robust classifier: (a) learning curves of the classification error and overall robust error over the 60 training epochs; (b) heatmap of the robust test error for pairwise class transformations based on the best trained classifier.

target transformation—the fraction of testing examples in class i that cannot be certified robust against transformation into class j for any ϵ norm-bounded attack. As shown in Figure 6.1(b), the vulnerability to adversarial transformations differs considerably among class pairs and appears correlated with perceptual similarity. For instance, only 0.26% of seeds in class 1 cannot be certified robust for target class 9 compare to 10% of seeds from class 9 into class 4.

Binary Cost Matrix. Next, we evaluate the effectiveness of cost-sensitive robustness training in producing models that are more robust for adversarial transformations designated as valuable. We consider four types of tasks defined by different binary cost matrices that capture different sets of adversarial transformations: *single pair*: particular seed class s to particular target class t; *single seed*: particular seed class s to any target class; *single target*: any seed class to particular target class t; and *multiple*: multiple seed and target classes. For each setting, the cost matrix is defined as $C_{ij} = 1$ if (i, j) is selected; otherwise, $C_{ij} = 0$. In general, we expect that the sparser the cost matrix, the more opportunity there is for

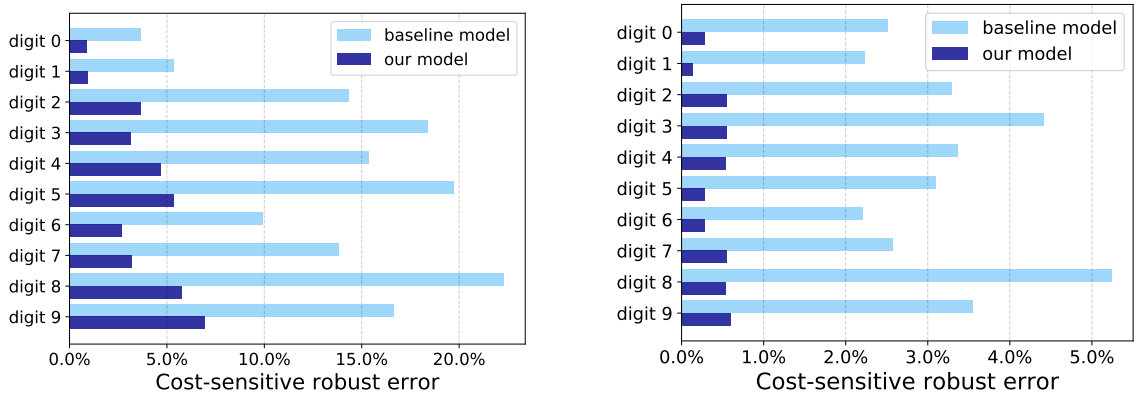
Table 6.1: Comparisons between different robust defense models on MNIST dataset against ℓ_∞ norm-bounded adversarial perturbations with $\epsilon = 0.2$. The sparsity gives the number of non-zero entries in the cost matrix over the total number of possible adversarial transformations. The candidates column is the number of potential seed examples for each task.

Task Description		Sparsity	Candidates	α	Standard Error		Robust Error	
					baseline	ours	baseline	ours
single pair	(0,2)	1/90	980	10.0	3.39%	2.68%	0.92%	0.31%
	(6,5)	1/90	958	5.0	3.39%	2.49%	3.55%	0.42%
	(4,9)	1/90	982	4.0	3.39%	3.00%	10.08%	1.02%
single seed	digit 0	9/90	980	10.0	3.39%	3.48%	3.67%	0.92%
	digit 2	9/90	1032	1.0	3.39%	2.91%	14.34%	3.68%
	digit 8	9/90	974	0.4	3.39%	3.37%	22.28%	5.75%
single target	digit 1	9/90	8865	4.0	3.39%	3.29%	2.23%	0.14%
	digit 5	9/90	9108	2.0	3.39%	3.24%	3.10%	0.29%
	digit 8	9/90	9026	1.0	3.39%	3.52%	5.24%	0.54%
multiple	top 10	10/90	6024	0.4	3.39%	3.34%	11.14%	7.02%
	random 10	10/90	7028	0.4	3.39%	3.18%	5.01%	2.18%
	odd digit	45/90	5074	0.2	3.39%	3.30%	14.45%	9.97%
	even digit	45/90	4926	0.1	3.39%	2.82%	13.13%	9.44%

cost-sensitive training to improve cost-sensitive robustness over models trained for overall robustness.

For the single pair task, we selected three representative adversarial goals: a low vulnerability pair (0, 2), medium vulnerability pair (6, 5) and high vulnerability pair (4, 9). We selected these pairs by considering the robust error results on the overall-robustness trained model (Figure 6.1(b)) as a rough measure for transformation hardness. This is generally consistent with intuitions about the MNIST digit classes (e.g., "9" and "4" look similar, so are harder to induce robustness against adversarial transformation), as well as with the visualization results produced by dimension reduction techniques, such as t-SNE [78].

Similarly, for the single seed and single target tasks we select three representative examples



(a) single seed class

(b) single target class

Figure 6.2: Cost-sensitive robust error using the proposed model and baseline model on MNIST for different binary tasks: (a) treat each digit as the seed class of concern respectively; (b) treat each digit as the target class of concern respectively.

representing low, medium, and high vulnerability to include in Table 6.1 and provide full results for all the single-seed and single target tasks for MNIST in Figure 6.2. For the multiple transformations task, we consider four variations: (i) the ten most vulnerable seed-target transformations; (ii) ten randomly-selected seed-target transformations; (iii) all the class transformations from odd digit seed to any other class; (iv) all the class transformations from even digit seed to any other class.

Table 6.1 summarizes the results, comparing the cost-sensitive robust error between the baseline model trained for overall robustness and a model trained using our cost-sensitive robust optimization. The cost-sensitive robust defense model is trained with $\epsilon = 0.2$ based on loss function (6.10) and the corresponding cost matrix $\mathbf{C}$. The regularization parameter α is tuned via cross validation. We report the selected best α, classification error and cost-sensitive robust error on the testing dataset.

Our model achieves a substantial improvement on the cost-sensitive robustness compared with the baseline model on all of the considered tasks, with no significant increases in normal classification error. The cost-sensitive robust error reduction varies from 30% to

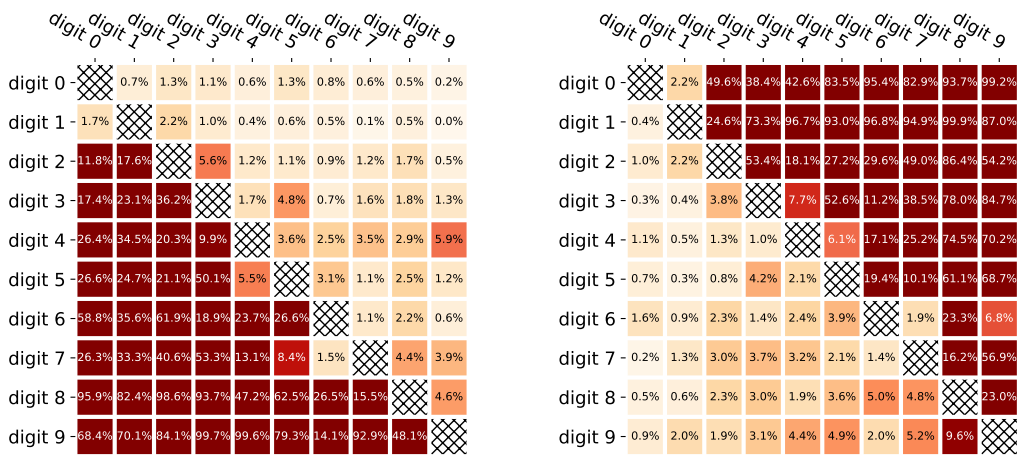
Table 6.2: Comparison results of different robust defense models for tasks with real-valued cost matrix.

Dataset	Task	Sparsity	Candidates	α	Standard Error		Robust Cost	
					baseline	ours	baseline	ours
MNIST	small-large	45/90	10000	0.04	3.39%	3.47%	2.245	0.947
MNIST	large-small	45/90	10000	0.04	3.39%	3.13%	3.344	1.549
CIFAR-10	vehicle	40/90	4000	0.1	31.80%	26.19%	4.183	3.095

90%, and is generally higher for sparse cost matrices. In particular, our classifier reduces the number of cost-sensitive adversarial examples from 198 to 12 on the single target task with digit 1 as the target class.

Real-valued Cost Matrices. Loosely motivated by a check forging adversary who obtains value by changing the semantic interpretation of a number [93], we consider two real-valued cost matrices: *small-large*, where only adversarial transformations from a smaller digit class to a larger one are valued, and the cost of valued-transformation is quadratic with the absolute difference between the seed and target class digits: $C_{ij} = (i-j)^2$ if $j > i$, otherwise $C_{ij} = 0$; *large-small*: only adversarial transformations from a larger digit class to a smaller one are valued: $C_{ij} = (i-j)^2$ if $i > j$, otherwise $C_{ij} = 0$. We tune α for the cost-sensitive robust model on the training MNIST dataset via cross validation, and set all the other parameters the same as in the binary case. The certified robust error for every adversarial transformation on MNIST testing dataset is shown in Figure 6.3, and the classification error and robust cost are given in Table 6.2. Compared with the model trained for overall robustness (Figure 6.1(b)), our trained classifier achieves stronger robustness guarantees on the adversarial transformations that induce costs, especially for those with larger costs.

Experiments on CIFAR-10.



(a) small-large (b) large-small

Figure 6.3: Heatmaps of robust test error using our cost-sensitive robust classifier on MNIST for various real-valued cost tasks: (a) *small-large*; (b) *large-small*.

We use the same neural network architecture for the CIFAR-10 dataset as [122], with four convolutional layers and two fully-connected layers. For memory and computational efficiency, we incorporate the approximation technique based on nonlinear random projection during the training phase ([122]). We train both the baseline model and our model using random projection of 50 dimensions, and optimize the training objective using SGD. Other parameters such as learning rate and batch size are set as same as those in [122].

Given a specific task, we train the cost-sensitive robust classifier on 80% randomly-selected training examples, and tune the regularization parameter α according to the performance on the remaining examples as validation dataset. The tasks are similar to those for MNIST (§6.2.3), except for the multiple transformations task we cluster the ten CIFAR-10 classes into two large groups: animals and vehicles, and consider the cases where only transformations between an animal class and a vehicle class are sensitive, and the converse.

Table 6.3 shows results on the testing data based on different robust defense models with $\epsilon = 2/255$. For all of the aforementioned tasks, our models substantially reduce the cost-

Table 6.3: Cost-sensitive robust models for CIFAR-10 dataset against adversarial examples, $\epsilon = 2/255$.

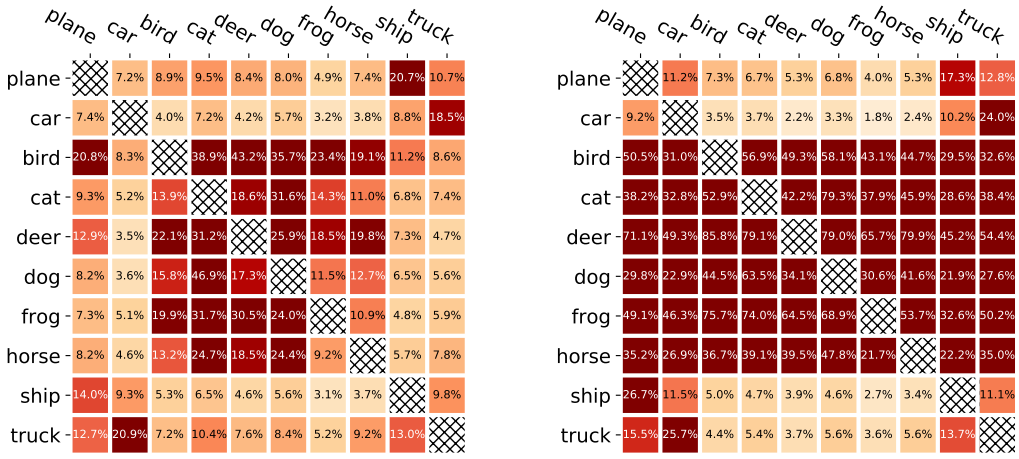
Task Description		Sparsity	Candidates	α	Standard Error		Robust Error	
					baseline	ours	baseline	ours
single pair	(frog, bird)	1/90	1000	10.0	31.80%	27.88%	19.90%	1.20%
	(cat, plane)	1/90	1000	10.0	31.80%	28.63%	9.30%	2.60%
single seed	dog	9/90	1000	0.2	31.80%	30.69%	57.20%	28.90%
	truck	9/90	1000	0.8	31.80%	31.55%	35.60%	15.40%
single target	deer	9/90	9000	0.1	31.80%	26.69%	16.99%	3.77%
	ship	9/90	9000	0.1	31.80%	24.80%	9.42%	3.06%
multiple	A-V	24/90	6000	0.1	31.80%	26.65%	16.67%	7.42%
	V-A	24/90	4000	0.2	31.80%	27.60%	12.07%	8.00%

sensitive robust error while keeping a lower classification error than the baseline.

For the real-valued task, we are concerned with adversarial transformations from seed examples in vehicle classes to other target classes. In addition, more cost is placed on transformations from vehicle to animal, which is 10 times larger compared with that from vehicle to vehicle. Figures 6.4(a) and 6.4(b) illustrate the pairwise robust test error using overall robust model and the proposed classifier for the aforementioned real-valued task on CIFAR-10.

Varying Adversary Strength. We investigate the performance of our model against different levels of adversarial strength by varying the value of ϵ that defines the ℓ_∞ ball available to the adversary. Figure 6.5 show the overall classification and cost-sensitive robust error of our best trained model, compared with the baseline model, on the MNIST single seed task with digit 9 and CIFAR-10 single seed task with dog as the seed class of concern, as we vary the maximum ℓ_∞ perturbation distance.

Under all the considered attack models, the proposed classifier achieves better cost-sensitive adversarial robustness than the baseline, while maintaining similar classification accuracy



(a) baseline model (b) our model

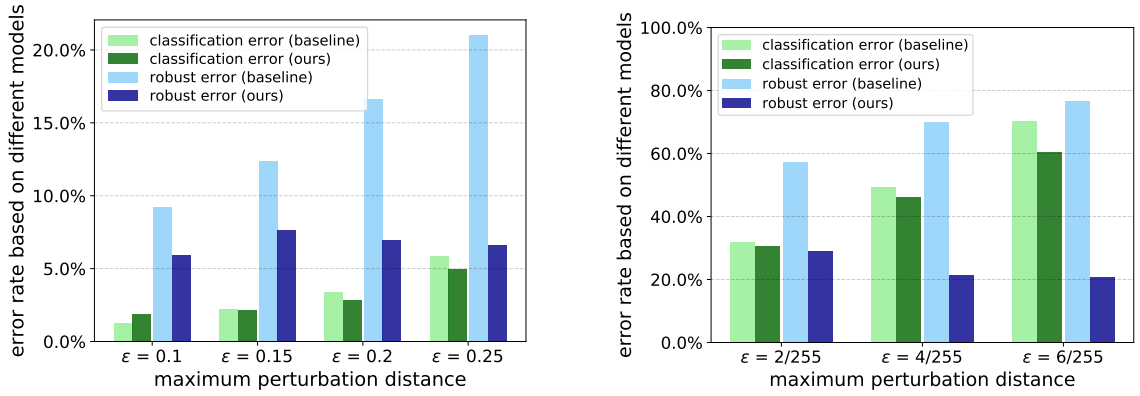
Figure 6.4: Heatmaps of robust test error for the real-valued task on CIFAR-10 using different robust classifiers: (a) baseline model; (b) our proposed cost-sensitive robust model.

on original data points. As the adversarial strength increases, the improvement for cost-sensitive robustness over overall robustness becomes more significant.

6.3 Uncertainty-Aware Robustness[4]

According to our experiments on CIFAR-10H, we know that ambiguous inputs, which are inherently difficult to assign a deterministic label even for human annotators, exist in benchmark image classification datasets such as CIFAR-10 (see Figure 4.2 in Chapter 4 for an illustration). Pervasive label errors have also been found in other most commonly-used computer vision, natural language, and audio datasets [90]. The existence of those examples will largely affect the intrinsic limits of achievable robustness that come from concentration of measure, as error regions of classifiers produced by state-of-the-art learning methods are likely to be around those examples. To escape such intrinsic barriers for building better robust machine learning systems, there is a need to rethink how to treat those inputs properly

[4]This is a working paper. We show preliminary results in this chapter.



(a) MNIST (b) CIFAR-10

Figure 6.5: Results for different adversary strengths, ϵ, for different settings: (a) MNIST single seed task with digit 9 as the chosen class; (b) CIFAR-10 single seed task with dog as the chosen class.

in the current learning framework for building robust models against adversarial examples.

An underlying assumption of overall robustness (Equation (6.1)) is that small perturbations preserve the ground-truth. For typical classification tasks in computer vision and natural language processing, the 'ground-truth' label of an input is usually assigned as the majority vote of human annotators. Although assigning a single label by majority vote simplifies the classification task, it removes the underlying probabilistic label information, especially for inputs with inherently uncertain labels. In this section, we argue that overall robustness would be an unrealistic goal with respect to uncertain inputs, thus should be modified to take into account the heterogeneity of such uncertainty.

6.3.1 Defining Uncertainty-Aware Robustness

In this section, we explain why the standard notion of overall robustness is not a good metric for assessing classifier's robustness property at uncertain inputs, and introduce the proposed notion of uncertainty-aware adversarial robustness.

We work with the same definition of label distribution as introduced in Chapter 4. Let $(\mathcal{X}, \mu)$ be the input probability space and $\mathcal{Y} = \{1, 2, \ldots, k\}$ be the set of possible labels. A function $\eta : \mathcal{X} \to [0,1]^k$ is said to capture the *full label distribution* [47, 45], if $[\eta(\boldsymbol{x})]_y$ represents the description degree of y to $\boldsymbol{x}$ for any $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, and $\sum_{y \in [k]} [\eta(\boldsymbol{x})]_y = 1$ for any $\boldsymbol{x} \in \mathcal{X}$. In probability theory, η can be understood as capturing the conditional probability distribution of $Y|X$. For classification tasks, the underlying concept function $c : \mathcal{X} \to \mathcal{Y}$ that gives 'ground-truth' label can be regarded as the Bayes optimal classifier with respect to η, namely $c(\boldsymbol{x}) = \text{argmax}_{y \in [k]} [\eta(\boldsymbol{x})]_y$.

Maximizing overall robustness as defined in (6.1) is typically regarded as the design goal for robust machine learning systems. Note that according to the definition, the concept function is assumed to assign a single label to each input as the ground-truth. However, for uncertain inputs whose class is intrinsically hard to determine, defining adversarial examples with reference to the underlying ground-truth with a single label would become controversial. As an example, if a CIFAR image looks like a dog to 60% of humans, and like a cat to the remaining 40%, it would not be reasonable to regard class transformations with target class of cat or dog as adversarial. Therefore, we propose the following definition to adapt the overall robustness to account for uncertain inputs, which are inspired by the literature on conformal classification [102, 3].

Definition 6.1 (Uncertainty-Aware Robustness)**.** Consider a probability space of inputs $(\mathcal{X}, \mu)$. Given parameter $\alpha \in [0,1]$ representing the threshold for constructing ground-truth label sets from η, the *uncertainty-aware robustness* of any classifier $f : \mathcal{X} \to \mathcal{Y}$ with respect to ϵ-perturbations bounded by metric Δ is defined as:

$$\widetilde{\text{AdvRob}}_\epsilon(f; \alpha) = 1 - \Pr_{\boldsymbol{x} \sim \mu} \left[\exists\, \boldsymbol{x}' \in \mathcal{B}_\epsilon(\boldsymbol{x}) \text{ s.t. } f(\boldsymbol{x}') \notin \mathcal{T}_\alpha(\boldsymbol{x}; \eta) \right],$$

where $\mathcal{T}_\alpha : \mathcal{X} \to \text{Pow}(\mathcal{Y})$ is a set-valued function that maps each input to a set of most probable labels based on η and the thresholding parameter α.

In Definition 6.1, $\mathcal{T}_\alpha(\boldsymbol{x}; \eta)$ can be understood as a set of plausible classes that $\boldsymbol{x}$ could be assigned to. For simplicity, we write $\mathcal{T}_\alpha(\boldsymbol{x}; \eta) = \mathcal{T}_\alpha(\boldsymbol{x})$ in the following discussions. One naive choice of the set-valued operator $\mathcal{T}$ is:

$$\mathcal{T}_\alpha(\boldsymbol{x}) = \big\{y \in \mathcal{Y} : [\eta(\boldsymbol{x})]_y \geq \alpha\big\}.$$

However, it has been shown in conformal classification literature that although such choice of $\mathcal{T}_\alpha$ produces smallest average set size [103], it tends to undercover hard subgroups while overcover easy ones [2]. To generate more adaptive prediction sets, the following construction rule of $\mathcal{T}_\alpha$ has been proposed:

$$\mathcal{T}_\alpha(\boldsymbol{x}) = \{\pi_1, \pi_2, \ldots, \pi_{k'}\}, \quad \text{where } k' = \inf\big\{l : \sum_{j=1}^{l} [\eta(\boldsymbol{x})]_{\pi_j} \geq \alpha\big\}, \tag{6.11}$$

and π is a permutation of $\{1, 2, \ldots, k\}$ that sorts $\eta(\boldsymbol{x})$ from most likely to least likely. Note that according to (6.11), $|\mathcal{T}_\alpha(\boldsymbol{x})|$ is monotonically non-decreasing with respect to α. In order to generate underlying ground-truth label sets that are adaptively to each individual input, we choose the latter construction rule for $\mathcal{T}_\alpha$ in the following discussions and experiments.

In contrast to the definition of overall robustness, only class transformations that do not belong to the underlying label set of $\mathcal{T}_\alpha(\boldsymbol{x})$ are considered as adversarial when uncertainty-aware robustness are considered. We argue that this is a better design goal for robustness, because uncertain inputs with $|\mathcal{T}_\alpha(\boldsymbol{x})| \geq 2$ are intrinsically more difficult to classify correctly by the underlying ground-truth, thus instead of treating those examples equally, we add more tolerance to perturbations with target class $y_{\text{targ}} \in \mathcal{T}_\alpha(\boldsymbol{x})$, which can be regarded

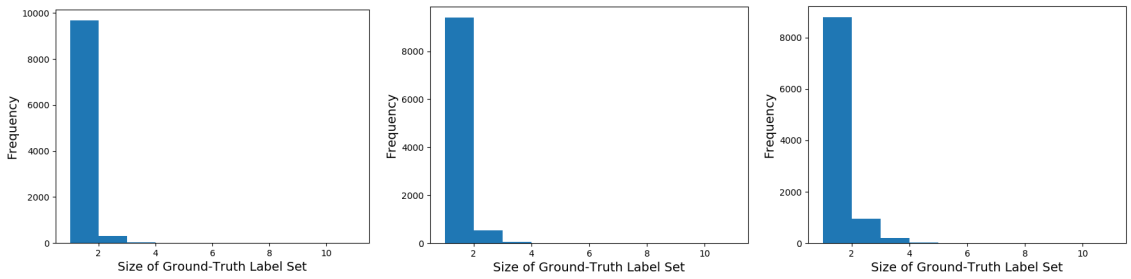
as relatively benign perturbations according to $\eta(\boldsymbol{x})$.

6.3.2 Measuring Uncertainty-Aware Robustness

In this section, we first discuss the challenges for empirically measuring the proposed notion of uncertainty-aware robustness, then propose our solutions to address such challenges.

To empirically evaluate the uncertainty-aware robustness of a given classifier, there exist two main challenges. First, the proposed definition requires knowledge of the underlying label distribution for any testing input, which is unavailable in typical benchmack datasets. As discussed in Section 4.3, human soft labels collected from multiple human annotators provide an approximate of such label distribution information for classification datasets that rely on human labeling. In our preliminary experiments, we make use of the CIFAR-10H dataset [95] which consists of human soft labels for CIFAR-10 test examples. If the human soft labels are unavailable for the dataset, an alternative solution which we are going to study as future work is to predict the underlying label set by adapting state-of-the-art conformal classification techniques [102, 3]. However, since conformal classifiers are typically designed based on some pretrained machine learning classifier, it remains an open question of whether the label set returned by a conformal classification procedure are able to approximate the ground-truth label distribution well.

Second, although we can use the same techniques for measuring overall robustness, such as attack-based optimization methods, to empirically measure the uncertainty-aware robustness with respect to inputs with $|\mathcal{T}_\alpha(\boldsymbol{x})| = 1$, it remains unclear how to deal with uncertain inputs with $|\mathcal{T}_\alpha(\boldsymbol{x})| \geq 2$. Noticing that as long as we can devise a attack with any possible target class $y_{\text{targ}} \notin \mathcal{T}_\alpha(\boldsymbol{x})$, we immediate know there exists an adversarial example by the definition of uncertainty-aware robustness. Therefore, we propose to search for the worst-



(a) $\alpha = 0.7$ (b) $\alpha = 0.8$ (c) $\alpha = 0.9$

Figure 6.6: Size distribution of ground-truth label set constructed based on (6.11) with $\alpha \in \{0.7, 0.8, 0.9\}$ for CIFAR-10 test dataset.

case perturbation for each possible adversarial class transformation based on the following objective:

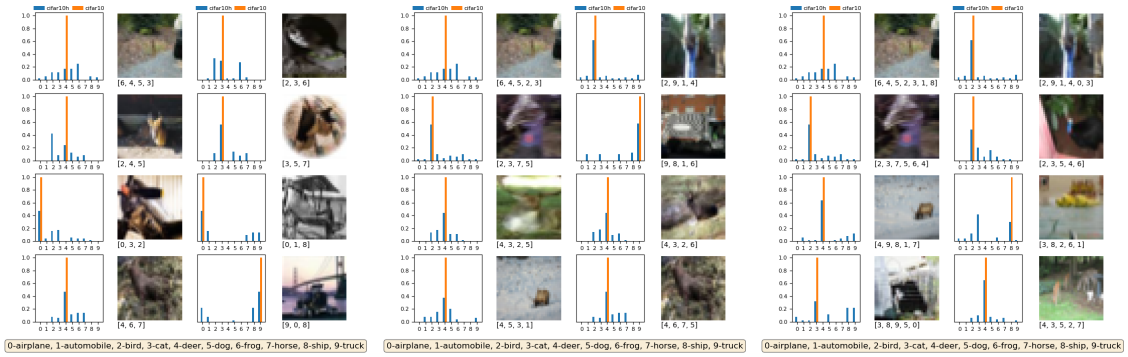
$$\underset{\boldsymbol{\delta} \in \mathcal{B}_\epsilon(\boldsymbol{x})}{\text{minimize}} \; \mathcal{L}\big(f_\theta(\boldsymbol{x} + \boldsymbol{\delta}), y_{\text{targ}}\big), \; \text{for any } y_{\text{targ}} \notin \mathcal{T}_\alpha(\boldsymbol{x}), \tag{6.12}$$

where f_θ is a given classifier parameterized by θ that is being evaluated and $\mathcal{B}_\epsilon(\boldsymbol{x})$ is the allowable perturbations with strength ϵ. To produce an approximated solution to equation (6.12), one can set $\mathcal{L}$ as a function such as cross-entropy loss and adapt the standard PGD-attack [79], which is adopted in our preliminary experiments. Studying and developing optimization techniques that better solve (6.12) can provide us better sense of security under the proposed definition of robustness, which would be an interesting future work.

6.3.3 Experiments

In this section, we report on experiments on CIFAR-10 [68] to evaluate the uncertainty-aware robustness of state-of-the-art adversarially-trained classifiers. We make use of the CIFAR-10H human soft labels [95] to approximate the underlying label distribution.

First, we visualize the generated ground-truth label sets using the construction rule of $\mathcal{T}_\alpha(\boldsymbol{x})$



(a) $\alpha = 0.7$ (b) $\alpha = 0.8$ (c) $\alpha = 0.9$

Figure 6.7: Visualizations of the top CIFAR-10 test images sorted by $|\mathcal{T}_\alpha(\boldsymbol{x})|$ with $\mathcal{T}$ constructed by (6.11) with α chosen from $\{0.7, 0.8, 0.9\}$.

defined in (6.11). Figure 6.6 illustrates the size distribution of the constructed ground-truth label sets on CIFAR-10. Note that larger label set size suggests that the image is intrinsically hard to classify. Most of the CIFAR-10 test images have $|\mathcal{T}_\alpha(\boldsymbol{x})| = 1$, which means all the human annotators agree on the label class that $\boldsymbol{x}$ should belong to. The set size distribution becomes more right-skewed as we increase α, which is aligned with the construction rule of (6.11). This is further demonstrated by Figure 6.7, which visualizes the top uncertain images that have the largest label set size. These images are indeed intrinsically hard to assign a deterministic label and the constructed label set captures the underlying label information for each image, which justifies the usefulness of the proposed construction rule of $\mathcal{T}_\alpha$. We also note that a few CIFAR-10 test images has originally assigned label not in our constructed label set $\mathcal{T}_\alpha(\boldsymbol{x})$ (see Figure 6.8 for examples). By visually examining these images, we can easily conclude that the original CIFAR-10 labels are incorrect. This further suggests that the standard notion of overall robustness will not be be a useful evaluation metric for such mislabeled examples.

Next, we evaluate the uncertainty-aware robustness of existing state-of-the-art adversarially-trained classifiers on CIFAR-10. We consider the most popular ℓ_∞-perturbations with

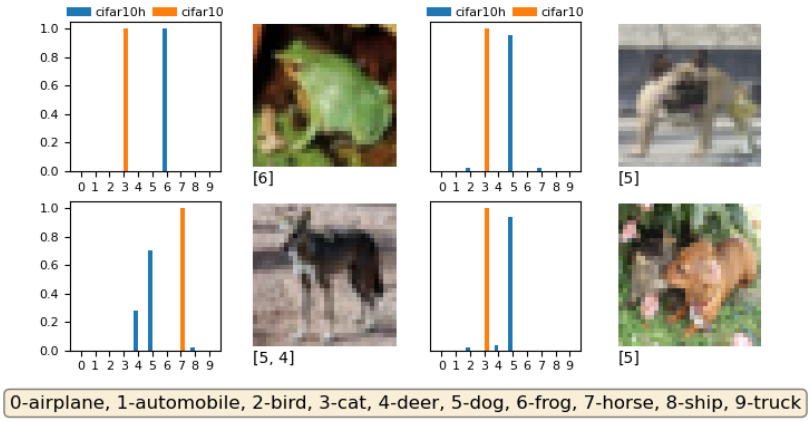


Figure 6.8: Illustration of CIFAR-10 images with original assigned label not covered by the constructed ground-truth label set $\mathcal{T}_\alpha(\boldsymbol{x})$ with $\alpha = 0.9$ based on the CIFAR-10H dataset.

strength $\epsilon = 8/255$. To implement the proposed method for measuring uncertainty-aware robustness for a given classifier, we adopt the 10-step PGD-attack with attack step size $\epsilon/4$ and three random restarts for solving the optimization problem (6.12). We set the thresholding parameter $\alpha = 0.9$ for generating ground-truth label set in this experiment. Table 6.4 illustrates the uncertainty-aware robustness, denoted as **Rob. Acc. (ls)**, of the considered classifier with respect to different groups of examples stratified by label set size. For comparisons, we also demonstrate the standard accuracy (**Std. Acc.**) and the overall robustness (**Rob. Acc.**) evaluated using standard PGD-attack with respect the original CIFAR-10 labels, and the uncertainty-aware standard accuracy (**Std. Acc. (ls)**) which is computed by setting $\epsilon = 0$ in Definition 6.1.

As we increase the set size threshold from 1 to 5, the overall robustness of the classifier decreases dramatically. This suggests that images with larger label sets are indeed more difficult to classify correctly, thus should be treated differently compared with images with a single deterministic label. When we shift our evaluation criterion of robustness from overall robustness to uncertainty-aware robustness, adversarial perturbations for uncertain inputs become harder to find, which is expected because adversaries are having less free-

Table 6.4: Evaluations of the uncertainty-aware robustness of the state-of-the-art adversairally-trained classifier [17] on group of CIFAR-10 test images with different intrinsic uncertainty level. Due to the randomness of the attack, we report both the mean estimate and its standard deviation over 3 repeated trials for Rob. Acc. and Rob. acc. (ls).

Group	**#Examples**	**Std. Acc.**	**Std. Acc. (ls)**	**Rob. Acc.**	**Rob. Acc. (ls)**
All	$10,000$	89.69%	92.52%	$63.86 \pm 0.02\%$	$68.88 \pm 0.02\%$
Size == 1	$8,789$	92.55%	92.56%	$68.69 \pm 0.01\%$	$68.81 \pm 0.01\%$
Size >= 2	$1,211$	68.95%	92.24%	$28.24 \pm 0.07\%$	$69.58 \pm 0.08\%$
Size >= 3	255	60.78%	92.16%	$17.39 \pm 0.18\%$	$71.37 \pm 0.32\%$
Size >= 4	60	63.33%	86.67%	$16.67 \pm 0.00\%$	$68.33 \pm 0.00\%$
Size >= 5	19	73.68%	78.95%	$10.53 \pm 0.00\%$	$52.63 \pm 0.00\%$

dom in selecting target vulnerable class. For instance, when we evaluate the set of images with $|\mathcal{T}_\alpha(\boldsymbol{x})| \geq 2$, the overall robustness is around 30% lower than the uncertainty-aware robustness. Similar trends was observed for other state-of-the-art classifiers as well as other ℓ_2-norm bounded perturbations (see Table 6.5).

The large difference on robustness performance with respect to the set of uncertain inputs suggests that a possible way to build better robust systems is to enable the learning method to account for such uncertainty information. Developing efficient and effective methods to train classifiers for uncertainty-aware robustness would be an important next step. However, unlike the testing CIFAR-10 images, we do not have the underlying label distribution information for training images. In the future, we are going to study how to adapt adversarial training [79, 129] and conformal classification methods [3] to address this challenge and train for uncertainty-aware robustness. We are hoping that studying uncertainty-aware robustness would be an initial step towards escaping the current limits and building better robust machine learning systems.

Table 6.5: Evaluations of the uncertainty-aware robustness of various state-of-the-art adversarially-trained classifiers with respect to the set of uncertain inputs with $|\mathcal{T}_\alpha(\boldsymbol{x})| \geq 2$ on CIFAR-10. We set the thresholding parameter $\alpha = 0.9$.

Metric	Strength	Model	Std. Acc.	Std. Acc. (ls)	Rob. Acc.	Rob. Acc. (ls)
ℓ_∞-norm	$\epsilon = 8/255$	[17]	68.95%	92.24%	$28.24 \pm 0.07\%$	$69.58 \pm 0.08\%$
		[123]	63.91%	89.93%	$26.97 \pm 0.04\%$	$64.38 \pm 0.10\%$
		[129]	62.18%	89.76%	$25.97 \pm 0.04\%$	$60.45 \pm 0.00\%$
ℓ_2-norm	$\epsilon = 0.5$	[7]	72.25%	88.19%	$43.96 \pm 0.08\%$	$62.07 \pm 0.10\%$
		[101]	69.03%	93.15%	$37.60 \pm 0.04\%$	$77.46 \pm 0.00\%$
		[123]	68.62%	90.67%	$45.69 \pm 0.08\%$	$80.10 \pm 0.00\%$

6.4 Summary

By focusing on overall robustness, previous robustness training methods expend a large fraction of the capacity of the network on unimportant transformations. In Section 6.2, we argue that for most scenarios, the actual harm caused by an adversarial transformation often varies depending on the seed and target class, so robust training methods should be designed to account for these differences. In Section 6.3, we argue that overall robustness is not properly defined with respect to intrinsically uncertain inputs, which should be treated in a different way compared with inputs that have deterministic label. We hope that considering cost-sensitive robustness and uncertainty-aware robustness instead of overall robustness will be an important step towards achieving more realistic, but still meaningful, robustness goals.

Chapter 7

Conclusion

In this dissertation, we developed an empirical framework to understand and estimate the intrinsic limits of adversarial robustness. The proposed framework connects theoretical works that prove the inevitability of adversarial examples [49, 44, 80, 108] with empirical studies that propose defenses against adversarial examples. We show that benchmark image datasets are *not* concentrated under typical perturbation metrics, thus unlike the impossibility results concluded in theoretical studies, standard concentration of measure can only explain a fairly small amount of adversarial vulnerability of state-of-the-art machine learning classifiers. Observing that labels are not considered in the standard concentration problem, we further investigate the usefulness of labels in defining more meaningful intrinsic robustness limits. We found that in addition to concentration of measure, the existence of uncertain inputs is another fundamental cause of intrinsic adversarial vulnerability, which are usually overlooked in adversarial machine learning literature.

Although the proposed framework is able to identify several intrinsic causes of adversarial vulnerability and provide quantifiable estimates of their effects, the intrinsic robustness limits that come from these discovered causes are still much higher than the best robustness performance achieved by state-of-the-art robustly-trained classifiers (see Figure 4.1 for an illustration). One possible explanation is that the concentration of measure problem (even with the label uncertainty constraint) is much simpler than the underlying problem of learning with adversarial examples, since the optimal solution to the concentration problem may

not be realized as error region of any learnable classifier. As a result, it is likely that intrinsic adversarial vulnerability could also originate from the underlying learning procedure of robust classifiers. It remains an open question whether it is plausible to characterize the intrinsic robustness limits with respect to the set of *learnable classifiers*, which would be the most realistic goal of robustness for empirical defenses to target for. Nevertheless, there remain challenges such as how to rigorously define the set of learnable classifiers and even with a definition of learnable classifiers, whether it is possible to be incorporated into the proposed concentration estimation framework. Perhaps, studying the intrinsic robustness limit with respect to classifiers that are produced by some specific learning methods could be an initial step towards characterizing the intrinsic robustness with learnable classifiers. Another open question is whether we can use the discovered fundamental causes of adversarial vulnerability to improve model robustness, potentially approaching the intrinsic robustness limit for the underlying task. A promising direction for future work is to study whether we could induce the classifier's error regions to reside in less concentrated region.

Moreover, we identify scenarios where overall robustness is not the right design goal for robust machine learning systems. We propose cost-sensitive robustness to take into account the potential harm of different adversarial class transformations; whereas we advocate for uncertainty-aware robustness as a better evaluation metric for inherently uncertain inputs. We design tools to train for cost-sensitive robustness, propose optimization methods to evaluate uncertainty-aware robustness, and demonstrate the usefulness of the proposed robustness definitions. By rethinking the current design goal of robustness, our works shed lights on potential ways to escape the intrinsic limits of robustness and build better robust machine learning systems. An open question on uncertainty-aware robustness is how to design a meaningful robustness metric if the classifiers are allowed output a set of labels. Literature on multi-label classification may provide important insights for this.

Bibliography

[1] Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? In *Advances in Neural Information Processing Systems*, pages 12192–12202, 2019.

[2] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[3] Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

[4] Kaiser Asif, Wei Xing, Sima Behpour, and Brian D Ziebart. Adversarial cost-sensitive classification. In *31st Conference on Uncertainty in Artificial Intelligence*, 2015.

[5] 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.

[6] Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, 2019.

[7] Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in-and out-distribution improves explainability. In *European Conference on Computer Vision*, pages 228–245. Springer, 2020.

[8] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, R Devon Hjelm, and Aaron C Courville. Mutual information neural estimation. In *International Conference on Learning Representations*, 2018.

[9] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[10] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, 2019.

[11] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[12] Christer Borell. The brunn-minkowski inequality in gauss space. *Inventiones mathematicae*, 30(2):207–216, 1975.

[13] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

[14] Sebastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pages 831–840, 2019.

[15] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, 2018.

[16] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

[17] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11190–11201, 2019.

[18] Thierry Champion, Luigi De Pascale, and Petri Juutinen. The ∞-wasserstein distance: Local solutions and existence of optimal transport maps. *SIAM Journal on Mathematical Analysis*, 40(1):1–20, 2008.

[19] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320, 2019.

[20] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[21] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[22] Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. PAC-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, 2018.

[23] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. Adversarial classification. In *10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2004.

[24] Georges A Darbellay and Igor Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

[25] Akshay Degwekar, Preetum Nakkiran, and Vinod Vaikuntanathan. Computational limitations in robust classification and win-win results. In *Conference on Learning Theory*, pages 994–1028, 2019.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[28] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer Science & Business Media, 2013.

[29] Guneet S. Dhillon, Kamyar Azizzadenesheli, Jeremy D. Bernstein, Jean Kossaifi, Aran Khanna, Zachary C. Lipton, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *International Conference on Learning Representations*, 2018.

[30] Dimitrios Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. Adversarial risk and robustness: General definitions and implications for the uniform distribution. In *Advances in Neural Information Processing Systems*, 2018.

[31] Elvis Dohmatob. Generalized no free lunch theorem for adversarial robustness. In *International Conference on Machine Learning*, pages 1646–1654, 2019.

[32] Pedro Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999.

[33] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, IV. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

[34] Tommaso Dreossi, Somesh Jha, and Sanjit A Seshia. Semantic adversarial deep learning. In *International Conference on Computer Aided Verification*, 2018.

[35] Krishnamurthy Dvijotham, Sven Gowal, Robert Stanforth, Relja Arandjelovic, Brendan O'Donoghue, Jonathan Uesato, and Pushmeet Kohli. Training verified learners with learned verifiers. *arXiv preprint arXiv:1805.10265*, 2018.

[36] David Eisenstat and Dana Angluin. The VC dimension of k-fold union. *Information Processing Letters*, 101(5):181–184, 2007.

[37] Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, 2001.

[38] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019.

[39] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Mądry. Adversarial robustness as a prior for learned representations. *arXiv:1906.00945*, 2019.

[40] Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.

[41] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola Schoenlieb. On the connection between adversarial robustness and saliency map interpretability. In *International Conference on Machine Learning*, pages 1823–1832, 2019.

[42] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[43] Kevin Eykholt, Swati Gupta, Atul Prakash, and Haizhong Zheng. Robust classification using robust feature augmentation. *arXiv:1905.10904*, 2019.

[44] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. In *Advances in Neural Information Processing Systems*, 2018.

[45] 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.

[46] Shivam Garg, Vatsal Sharan, Brian Zhang, and Gregory Valiant. A spectral view of adversarially robust features. In *Advances in Neural Information Processing Systems*, pages 10138–10148, 2018.

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

[48] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018.

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

[50] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

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

[52] 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 (ICCV)*, 2019.

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

[54] John A Hartigan and Manchek A Wong. A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[55] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[56] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, 2017.

[57] Julien M Hendrickx and Alex Olshevsky. Matrix p-norms are NP-hard to approximate if $p \neq 1, 2, \infty$. *SIAM Journal on Matrix Analysis and Applications*, 31(5), 2010.

[58] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.

[59] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[60] Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. O2U-Net: A simple noisy label detection approach for deep neural networks. In *International Conference on Computer Vision*, 2019.

[61] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.

[62] Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: analysis and improvement. In *Computer Vision and Pattern Recognition*, 2018.

[63] Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, et al. Nonparametric von mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems*, pages 397–405, 2015.

[64] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3573–3587, 2018.

[65] Justin Khim and Po-Ling Loh. Adversarial risk bounds for binary classification via function transformation. *arXiv preprint arXiv:1810.09519*, 2, 2018.

[66] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[67] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

[68] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[69] Ryen Krusinga, Sohil Shah, Matthias Zwicker, Tom Goldstein, and David Jacobs. Understanding the (un)interpretability of natural image distributions using generative models. *arXiv preprint arXiv:1901.01499*, 2019.

[70] Matjaž Kukar and Igor Kononenko. Cost-sensitive learning with neural networks. In *13th European Conference on Artificial Intelligence*, 1998.

[71] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017.

[72] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[73] Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. `http://yann.lecun.com/exdb/mnist`, 2010.

[74] Paul Lévy. *Problèmes concrets d'analyse fonctionnelle*, volume 6. Gauthier-Villars Paris, 1951.

[75] Ralph Linsker. How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Computation*, 1(3):402–411, 1989.

[76] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, 2018.

[77] Xu-Ying Liu and Zhi-Hua Zhou. The influence of class imbalance on cost-sensitive learning: An empirical study. In *Sixth International Conference on Data Mining*, 2006.

[78] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008.

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

[80] Saeed Mahloujifar, Dimitrios I 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*, 2019.

[81] Saeed Mahloujifar, Xiao Zhang, Mohammad Mahmoody, and David Evans. Empirically measuring concentration: Fundamental limits on intrinsic robustness. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[82] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pages 3578–3586, 2018.

[83] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[84] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2018.

[85] Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. *Proceedings of Machine Learning Research*, 99:1–19, 2019.

[86] Kevin R Moon, Kumar Sricharan, and Alfred O Hero. Ensemble estimation of mutual information. In *IEEE International Symposium on Information Theory*, pages 3030–3034. IEEE, 2017.

[87] Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318, 1995.

[88] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NeurIPS*, 2013.

[89] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, 2017.

[90] Curtis Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *NeurIPS (Datasets and Benchmarks Track)*, 2021.

[91] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

[92] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning (ICML)*, 2017.

[93] 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.

[94] Ankit Pensia, Varun Jog, and Po-Ling Loh. Extracting robust and accurate features via a robust information bottleneck. *IEEE Journal on Selected Areas in Information Theory*, 2020.

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

[96] 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.

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

[98] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

[99] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7909–7919, 2020.

[100] Mustapha Raıssouli and Iqbal H Jebril. Various proofs for the decrease monotonicity of the Schatten's power norm, various families of $\mathbb{R}^n$-norms and some open problems. *International Journal of Open Problems in Computer Science and Mathematics*, 3(2):164–174, 2010.

[101] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

[102] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

[103] Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

[104] Joshua Saxe and Konstantin Berlin. Deep neural network based malware detection using two dimensional binary program features. In *10th International Conference on Malicious and Unwanted Software*, 2015.

[105] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018.

[106] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on MNIST. In *International Conference on Learning Representations*, 2019.

[107] Clayton D Scott and Robert D Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7(Apr):665–704, 2006.

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

[109] Mahmood Sharif, Lujo Bauer, and Michael K Reiter. On the suitability of l_p-norms for creating and preventing adversarial examples. In *CVPR Workshop on Bright and Dark Sides of Computer Vision: Challenges and Opportunities for Privacy and Security*, 2018.

[110] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *ACM Conference on Computer and Communications Security*, pages 1528–1540, 2016.

[111] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[112] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[113] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[114] Vladimir N Sudakov and Boris S Tsirel'son. Extremal properties of half-spaces for spherically invariant measures. *Journal of Soviet Mathematics*, 9(1):9–18, 1978.

[115] Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In *New challenges for feature selection in data mining and knowledge discovery*, pages 5–20, 2008.

[116] 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.

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

[118] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33:1633–1645, 2020.

[119] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[120] Shiqi Wang, Yizheng Chen, Ahmed Abdou, and Suman Jana. MixTrain: Scalable training of formally robust neural networks. *arXiv preprint arXiv:1811.02625*, 2018.

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

[122] Eric Wong, Frank R Schmidt, Jan Hendrik Metzen, and Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, 2018.

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

[124] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018.

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

[126] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

[127] Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, 2019.

[128] Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE International Conference on Data Mining*, 2003.

[129] 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*, pages 7472–7482, 2019.

[130] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. *arXiv preprint arXiv:1906.06316*, 2019.

[131] Xiao Zhang, Jinghui Chen, Quanquan Gu, and David Evans. Understanding the intrinsic robustness of image distributions using conditional generative models. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3883–3893. PMLR, 26–28 Aug 2020.

[132] Xiao Zhang and David Evans. Cost-sensitive robustness against adversarial examples. In *International Conference on Learning Representations*, 2019.

[133] Xiao Zhang and David Evans. Understanding intrinsic robustness using label uncertainty. In *International Conference on Learning Representations*, 2022.

[134] Zhi-Hua Zhou and Xu-Ying Liu. On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257, 2010.

[135] Sicheng Zhu, Xiao Zhang, and David Evans. Learning adversarially robust representations via worst-case mutual information maximization. In *International Conference on Machine Learning*, pages 11609–11618. PMLR, 2020.