Empirically Measuring Concentration: Fundamental Limits on Intrinsic Robustness

Saeed Mahloujifar*, Xiao Zhang*, Mohammad Mahmoody, and David Evans
University of Virginia
[saeed, shawn, mohammad, evans]@virginia.edu

Abstract

Many recent works have shown that adversarial examples that fool classifiers can be found by minimally perturbing a normal input. Recent theoretical results, starting with Gilmer et al. (2018b), show that if the inputs are drawn from a *concentrated* metric probability space, then adversarial examples with small perturbation are inevitable. A concentrated space has the property that any subset with $\Omega(1)$ (e.g., 1/100) measure, according to the imposed distribution, has small distance to almost all (e.g., 99/100) of the points in the space. It is not clear, however, whether these theoretical results apply to actual distributions such as images. This paper presents a method for empirically measuring and bounding the concentration of a concrete dataset which is proven to converge to the actual concentration. We use it to empirically estimate the intrinsic robustness to ℓ_∞ and ℓ_2 perturbations of several image classification benchmarks. Code for our experiments is available at https://github.com/xiaozhanguva/Measure-Concentration.

1 Introduction

Despite achieving exceptionally high accuracy on natural inputs, state-of-the-art machine learning models have been shown to be vulnerable to adversaries who use small perturbations to fool the classifier (Szegedy et al., 2014; Goodfellow et al., 2015). This phenomenon, known as *adversarial examples*, has motivated numerous studies (Papernot et al., 2016; Madry et al., 2018; Biggio & Roli, 2018; Gilmer et al., 2018a) to develop heuristic defenses that aim to improve classifier robustness. However, most defense mechanisms have been quickly broken by adaptive attacks (Carlini & Wagner, 2017; Athalye et al., 2018). Although certification methods (Raghunathan et al., 2018; Wong & Kolter, 2018; Sinha et al., 2018; Wong et al., 2018; Gowal et al., 2019; Wang et al., 2018; Zhang et al., 2019) have been proposed aiming to end such arms race and continuous efforts have been made to develop better robust models, both the robustness guarantees and efficiency achieved by state-of-the-art robust classifiers are far from satisfying.

This motivates a fundamental information-theoretic question: *what are the inherent limitations of developing robust classifiers?* Several recent works (Gilmer et al., 2018b; Fawzi et al., 2018; Mahloujifar et al., 2019; Shafahi et al., 2019; Bhagoji et al., 2019) 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, Gilmer et al. (2018b) 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). Mahloujifar et al. (2019) generalized this result to any concentrated metric probability space of inputs and showed, for example, that if the inputs come from

*Equal contribution.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

any Normal Lévy family (Lévy, 1951), any classifier with a noticable test error will be vulnerable to small (i.e., sublinear in the typical norm of the inputs) 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. Our work develops a general method for testing properties of concrete datasets against these theoretical assumptions.

Contributions. Our work shrinks the gap between theoretical analyses of robustness of classification for theoretical data distributions and understanding the intrinsic robustness of actual datasets. Indeed, quantitative estimates of the intrinsic robustness[2] of benchmark image datasets such as MNIST and CIFAR-10 can provide us with a better understanding of the threat of adversarial examples for natural image distributions and may suggest promising directions for further improving classifier robustness. Our main technical contribution is a general method to evaluate the concentration of a given input distribution μ based on a set of data samples. We prove that by simultaneously increasing the sample size m and a complexity parameter T, the concentration of the empirical measure converges to the actual concentration of μ (Section 3). Using this method, we perform experiments to demonstrate the existence of robust error regions for benchmark datasets under both ℓ_∞ and ℓ_2 perturbations (Section 4). Compared with state-of-the-art robustly trained models, our estimated intrinsic robustness shows that, for most settings, there exists a large gap between the robust error achieved by the best current models and the theoretical limits implied by concentration. This suggests the concentration of measure is not the only reason behind the vulnerability of existing classifiers to adversarial perturbations. 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.

Related Work. We are aware of only one previous work that attempts to heuristically estimate these properties. To extend their theoretical impossibility result to the practical distributions, Gilmer et al. (2018b) studied MNIST dataset to find a region that is somewhat robust in terms of the *expected* ℓ_2 distance of other images from the region. In their setting, they showed the existence of a set of measure 0.01 with average ℓ_2 distance 6.59 to all points. In comparison, our work is the first to provide a general methodology to empirically estimate the concentration of measure with provable guarantees. Moreover, we are able to deal with ℓ_∞, and *worst-case* bounded perturbations for modeling adversarial risk, which is the most popular setting for research in adversarial examples. In addition, another related concurrent work (Bhagoji et al., 2019) studied lower bounds on the adversarial risk using optimal transport on the metric probability space of instances. They also measure the optimal transport on the empirical distributions but do not characterize the relationship between the optimal transport of empirical datasets and the actual one of the underlying distributions.

Another related line of work estimated lower bounds on the concentration of measure of the underlying distribution through simulating distributions by generative models. Fawzi et al. (2018) proved a lower bound on the concentration of the generated image distribution, assuming the underlying generative model has Gaussian latent space and small Lipschitz constant. Krusinga et al. (2019) estimated an upper bound on the density function of the distribution using generative model, then proved concentration inequalities based on upper bounds on the density function. Our work is distinct from these works, because we directly learn the concentration function instead of a lower bound, and we use the actual data samples instead of samples generated from some trained generative model.

The work of Tsipras et al. (2019) studied the trade-off between robustness and accuracy. They show that for some specific learning problems, achieving robustness and accuracy together is not possible. At first glance, it might seem that this trade-off contradicts the existing lower bounds 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 Diochnos et al. (2018)), but they coincide in the cases that the ground truth function is robust to small perturbations.

[2]See Definition 2.2 for the formal definition of intrinsic robustness. The term robustness has been used with different meanings in previous works (e.g., in Diochnos et al. (2018), it refers to the average distances to the error region). However, all such uses refer to a desirable property of the classifier in being resilient to adversarial perturbations, which is the case here as well. See Diochnos et al. (2018) for a taxonomy of different definitions.

Notation. Lowercase boldface letters such as $\boldsymbol{x}$ are used to denote vectors, and $[n]$ is used to represent $\{1, 2, \ldots, n\}$. For any set $\mathcal{A}$, let $\mathsf{Pow}(\mathcal{A})$, $|\mathcal{A}|$ and $\mathbb{1}_{\mathcal{A}}(\cdot)$ be the set of measurable subsets of $\mathcal{A}$, cardinality and indicator function of $\mathcal{A}$, respectively. For any $\boldsymbol{x} \in \mathbb{R}^n$, the ℓ_∞-norm and ℓ_2-norm of $\boldsymbol{x}$ are defined as $\|\boldsymbol{x}\|_\infty = \max_{i\in[n]} |x_i|$ and $\|\boldsymbol{x}\|_2 = (\sum_{i\in[n]} x_i^2)^{1/2}$ respectively. Let $(\mathcal{X}, \mu)$ be a probability space and $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be some distance metric defined on $\mathcal{X}$. Define the empirical measure with respect to a set $\mathcal{S}$ sampled from μ as $\hat{\mu}_{\mathcal{S}}(\mathcal{A}) = \sum_{\boldsymbol{x}\in\mathcal{S}} \mathbb{1}_{\mathcal{A}}(\boldsymbol{x})/|\mathcal{S}|, \forall \mathcal{A} \subseteq \mathcal{X}$. Let $\mathrm{Ball}(\boldsymbol{x}, \epsilon) = \{\boldsymbol{x}' \in \mathcal{X} : d(\boldsymbol{x}', \boldsymbol{x}) \leq \epsilon\}$ be the ball around $\boldsymbol{x}$ with radius ϵ. For any subset $\mathcal{A} \subseteq \mathcal{X}$, define the ϵ-expansion $\mathcal{A}_\epsilon = \{\boldsymbol{x} \in \mathcal{X} \colon \exists\, \boldsymbol{x}' \in \mathrm{Ball}(\boldsymbol{x}, \epsilon) \cap \mathcal{A}\}$. The collection of the ϵ-expansions for members of any $\mathcal{G} \subseteq \mathsf{Pow}(\mathcal{X})$ is defined and denoted as $\mathcal{G}_\epsilon = \{\mathcal{A}_\epsilon : \mathcal{A} \in \mathcal{G}\}$.

2 Robustness and Concentration of Measure

In this paper, we work with the following definition of *adversarial risk*:

Definition 2.1 (Adversarial Risk)**.** *Let $(\mathcal{X}, \mu)$ be the probability space of instances and f^* be the underlying ground-truth. The* adversarial risk *of a classifier f in metric d with strength ϵ is defined as*

$$\mathrm{AdvRisk}_\epsilon(f, f^*) = \Pr_{\boldsymbol{x}\leftarrow\mu}\left[\exists\, \boldsymbol{x}' \in \mathrm{Ball}(\boldsymbol{x}, \epsilon) \text{ s.t. } f(\boldsymbol{x}') \neq f^*(\boldsymbol{x}')\right].^3$$

For $\epsilon = 0$, which allows no perturbation, the notion of adversarial risk coincides with traditional risk.

Definition 2.2 (Intrinsic Robustness)**.** *Consider the same setting as in Definition 2.1. Let $\mathcal{F}$ be some family of classifiers, then the* intrinsic robustness *is defined as the maximum adversarial robustness that can be achieved within $\mathcal{F}$, namely*

$$\mathrm{Rob}_\epsilon(\mathcal{F}, f^*) = 1 - \inf_{f\in\mathcal{F}} \left\{\mathrm{AdvRisk}_\epsilon(f, f^*)\right\}.$$

In this work, we specify $\mathcal{F}$ as the family of imperfect classifiers that have risk at least $\alpha \in (0, 1)$.

Previous work shows a connection between concentration of measure and the intrinsic robustness with respect to some families of classifiers (Gilmer et al. (2018b); Fawzi et al. (2018); Mahloujifar et al. (2019); Shafahi et al. (2019)). The concentration of measure on a metric probability space is defined by a concentration function as follows.

Definition 2.3 (Concentration Function)**.** *Consider a metric probability space $(\mathcal{X}, \mu, d)$. Suppose $\epsilon > 0$ and $\alpha \in (0, 1)$ are given parameters, then the* concentration function *of the probability measure μ with respect to ϵ, α is defined as*

$$h(\mu, \alpha, \epsilon) = \inf_{\mathcal{E}\in\mathsf{Pow}(\mathcal{X})} \left\{\mu(\mathcal{E}_\epsilon) \colon \mu(\mathcal{E}) \geq \alpha\right\}.$$

Note that the standard notion of concentration function (e.g., see Talagrand (1995)) is related to a special case of Definition 2.3 by fixing $\alpha = 1/2$.

Generalizing the result of Gilmer et al. (2018b) about instances drawn from spheres, Mahloujifar et al. (2019) showed that, in general, if the metric probability space of instances is concentrated, then any classifier with 1% risk incurs large adversarial risk for small amount of perturbations.

Theorem 2.4 (Mahloujifar et al. (2019))**.** *Let $(\mathcal{X}, \mu)$ be the probability space of instances and f^* be the underlying ground-truth. For any classifier f, we have*

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

In order for this theorem to be useful, we need to know the concentration function. The behavior of this function is studied extensively for certain theoretical metric probability spaces (Ledoux, 2001; Milman & Schechtman, 1986). However, it is not known how to measure the concentration function for arbitrary metric probability spaces. In this work, we provide a framework to (algorithmically) bound the concentration function from i.i.d. samples from a distribution. Namely, we want to solve the following optimization task using our i.i.d. samples:

$$\underset{\mathcal{E}\in\mathsf{Pow}(\mathcal{X})}{\text{minimize}} \;\; \mu(\mathcal{E}_\epsilon) \quad \text{subject to } \mu(\mathcal{E}) \geq \alpha. \tag{1}$$

[3] Note that bounding l_p norm might be restrictive for the adversary (Gilmer et al., 2018a) and this definition only covers a subset of possible adversaries.

We aim to estimate the minimum possible adversarial risk, which captures the intrinsic robustness for classification in terms of the underlying distribution μ, conditioned on the fact that the original risk is at least α. Note that solving this optimization problem only shows the possibility of existence of an error region $\mathcal{E}$ with certain (small) expansion. This means that there could potentially exist a classifier with risk at least α and adversarial risk equal to the solution of the optimization problem of (1). Actually *finding* such an optimally robust classifier (with error α) using a learning algorithm might be a much more difficult task or even infeasible. We do not consider that problem in this work.

3 Method for Measuring Concentration

In this section, we present a method to measure the concentration of measure on a metric probability space using i.i.d. samples. To measure concentration, there are two main challenges:

1. Measuring concentration appears to require knowledge of the density function of the distribution, but we only have a data set sampled from the distribution.
2. Even with the density function, we have to find the best possible subset among all the subsets of the space, which seems infeasible.

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 3.1 (Concentration Function for a Collection of Subsets)**.** *Consider a metric probability space $(\mathcal{X}, \mu, d)$. Let $\epsilon > 0$ and $\alpha \in (0, 1)$ be given parameters, then the* concentration function *of the probability measure μ with respect to ϵ, α and a collection of subsets $\mathcal{G} \subseteq \mathsf{Pow}(\mathcal{X})$ is defined as*

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

When $\mathcal{G} = \mathsf{Pow}(\mathcal{X})$, we write $h(\mu, \alpha, \epsilon)$ for simplicity.

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 3.2 (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*

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

Theorem 3.3 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. The proof of Theorem 3.3 is provided in Appendix A.1.

Theorem 3.3 (Generalization of Concentration)**.** *Let $(\mathcal{X}, \mu, d)$ be a metric probability space and $\mathcal{G} \subseteq \mathsf{Pow}(\mathcal{X})$. For any $\delta, \alpha, \epsilon \in [0, 1]$, we have*

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

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

Remark 3.4. Theorem 3.3 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 3.3 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 Neyshabur et al. (2017)), 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 (Cullina et al., 2018; Attias et al., 2019; Montasser et al., 2019; Yin et al., 2019; Raghunathan et al., 2019).

The following theorem, proved in Appendix A.2, 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 3.5 and the techniques used in its proof are inspired by the work of Scott & Nowak (2006) on learning minimum volume sets.

Theorem 3.5. *Let $\{\mathcal{G}(T)\}_{T\in\mathbb{N}}$ be a family of subset collections defined over a space $\mathcal{X}$. Let $\left\{\phi^T\right\}_{T\in\mathbb{N}}$ and $\left\{\phi_\epsilon^T\right\}_{T\in\mathbb{N}}$ be two families of complexity penalty functions such that ϕ^T and ϕ_ϵ^T 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 $\lim_{T\to\infty} h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{G}(T)) = h(\mu, \alpha, \epsilon)$.

Remark 3.6. In Theorem 3.5, 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 $\phi_\epsilon^T(m(T), \delta(T))$ 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.

3.1 Special Case of ℓ_∞

In this subsection, we show how to instantiate Theorem 3.5 for the case of ℓ_∞. Below, we introduce a special collection of subsets characterized by the *complement of a union of hyperrectangles*:

Definition 3.7 (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 hyperrectangles *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:

$$\underset{\mathcal{E}\in\mathcal{CR}(T)}{\text{minimize}} \;\; \hat{\mu}_{\mathcal{S}}(\mathcal{E}_{\epsilon_\infty}) \quad \text{subject to } \hat{\mu}_{\mathcal{S}}(\mathcal{E}) \geq \alpha. \tag{2}$$

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 3.8. *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 3.5. In fact, we can tune the parameters more carefully to get T^2, instead of T^4, but the convergence will be slower. See Appendix A.3 for the proof.

3.2 Special Case of ℓ_2

This subsection demonstrates how to apply Theorem 3.5 to the case of ℓ_2. The following definition introduces the collection of subsets characterized by a *union of balls*:

Definition 3.9 (Union of Balls)**.** *For any positive integer T, the collection of subsets specified by a* union of T n-dimensional balls *is defined as*

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

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

$$\underset{\mathcal{E}\in\mathcal{B}(T)}{\text{minimize}} \;\; \hat{\mu}_{\mathcal{S}}(\mathcal{E}_{\epsilon_2}) \quad \text{subject to } \hat{\mu}_{\mathcal{S}}(\mathcal{E}) \geq \alpha. \tag{3}$$

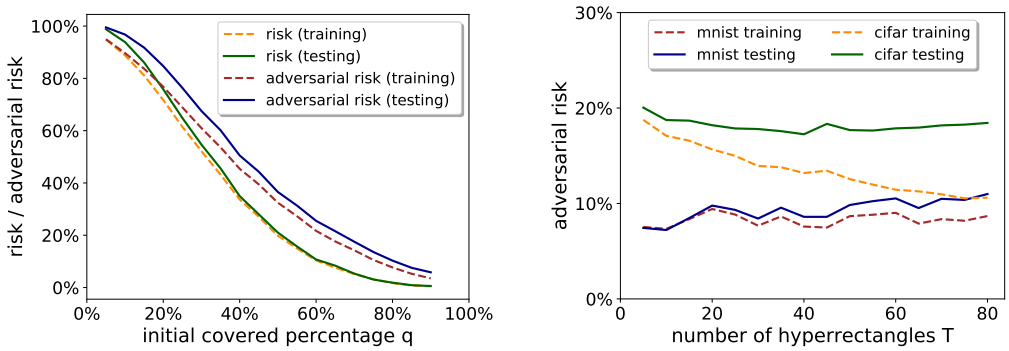
Theorem 3.10, proven in Appendix A.4, guarantees that if we increase the number of balls and samples together in a careful way, the solution to the empirical problem (3) converges to the true concentration.

Theorem 3.10. *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*

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

4 Experiments

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 both ℓ_∞ and ℓ_2 settings. Specifically, we first introduce our algorithm, then evaluate our approach on two benchmark image datasets: MNIST (LeCun et al., 2010) and CIFAR-10 (Krizhevsky & Hinton, 2009). Note that in our experiments we exactly use the collection of subsets as suggested by our theoretical results in the previous section. 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.



(a) varying q (b) varying T

Figure 1: (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 ($\epsilon_\infty = 8/255$).

4.1 Experiments for ℓ_∞

Theorem 3.8 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, to measure the concentration of μ, it remains to solve the optimization problem (2).

Method. Although the collection of subsets is specified using simple topology, solving (2) 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., $f_{\mathcal{E}}(\boldsymbol{x}) = f^*(\boldsymbol{x})$, if $\boldsymbol{x} \notin \mathcal{E}$; $f_{\mathcal{E}}(\boldsymbol{x}) \neq f^*(\boldsymbol{x})$, if $\boldsymbol{x} \in \mathcal{E}$. However, finding a classifier corresponding to $f_{\mathcal{E}}$ using a learning algorithm might be a very difficult task. Here, we find the optimally robust error region, not the corresponding classifier. A desirable error region should have small adversarial risk[4], 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 1 in Appendix B) 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 (Hartigan & Wong, 1979) 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 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 1). 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 1 shows the learning curves regarding risk and adversarial risk for two specific experimental settings (similar results are obtained under other experimental settings, see Appendix C.3). Figure 1(a) suggests that as we increase the initial covered percentage q, both risk and adversarial risk of the corresponding error region decrease. This supports our use of binary search on q in Algorithm 1. On

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

Table 1: Summary of the main results using our method for different settings with ℓ_∞ perturbations.

Dataset	α	ϵ_∞	T	**Best** q	**Empirical Risk** (%)		**Empirical AdvRisk** (%)	
					training	testing	training	testing
MNIST	0.01	0.1	5	0.662	1.22 ± 0.11	1.23 ± 0.12	3.65 ± 0.29	3.64 ± 0.30
		0.2	10	0.660	1.12 ± 0.13	1.11 ± 0.10	5.76 ± 0.38	5.89 ± 0.44
		0.3	10	0.629	1.12 ± 0.12	1.15 ± 0.13	7.34 ± 0.38	7.24 ± 0.38
		0.4	10	0.598	1.15 ± 0.09	1.21 ± 0.09	9.89 ± 0.57	9.92 ± 0.60
CIFAR-10	0.05	$2/255$	10	0.680	5.32 ± 0.21	5.72 ± 0.25	7.29 ± 0.20	8.13 ± 0.26
		$4/255$	20	0.688	5.59 ± 0.25	6.05 ± 0.40	11.43 ± 0.24	13.66 ± 0.33
		$8/255$	40	0.734	5.55 ± 0.21	5.94 ± 0.34	13.69 ± 0.19	18.13 ± 0.30
		$16/255$	75	0.719	5.16 ± 0.25	5.28 ± 0.23	19.77 ± 0.22	28.83 ± 0.46

Table 2: 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**	**Empirical Risk**	**Empirical AdvRisk**
MNIST	$\epsilon_\infty = 0.3$	Madry et al. (2018)	1.20%	10.70%
		Ours ($T = 10, \alpha = 0.012$)	$1.35\% \pm 0.08\%$	$8.28\% \pm 0.22\%$
MNIST	$\epsilon_2 = 1.5$	Schott et al. (2019)	1.00%	20.00%
		Ours ($T = 20, \alpha = 0.01$)	1.08%	2.12%
CIFAR-10	$\epsilon_\infty = 8/255$	Madry et al. (2018)	12.70%	52.96%
		Ours ($T = 40, \alpha = 0.127$)	$14.22\% \pm 0.46\%$	$29.21\% \pm 0.35\%$

the other hand, as can be seen from Figure 1(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 1 summarizes the optimal parameters, the empirical risk and adversarial risk of the corresponding error region on both training and testing datasets for each experimental setting (see Appendix C.1 for similar results on Fashion-MNIST and SVHN). 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 1 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-measured adversarial risk 52.96% (Madry et al., 2018).

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 Madry et al. (2018), 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 Madry et al. (2018), which is demonstrated in Table 2. 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 Madry et al. (2018). 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: Comparisons between different methods for finding robust error region with ℓ_2 perturbations.

Dataset	α	ϵ_2	Gilmer et al. (2018b)		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%

4.2 Experiments for ℓ_2

For ℓ_2 adversaries, Theorem 3.10 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). 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 2 in Appendix B) 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 (Gilmer et al., 2018b) on MNIST and CIFAR-10. In particular, 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 Wong et al. (2018). Table 3 summarizes the optimal parameters, the testing risk and adversarial risk (see Appendix C.2 for more detailed results, including for other datasets) 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 Gilmer et al. (2018b), 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 rates (see Table 2 for a comparison with the best robust classifier against ℓ_2 perturbations proposed in Schott et al. (2019)).

5 Conclusion

To understand whether theoretical results showing limits of intrinsic robustness for natural distributions apply to concrete datasets, we developed a general framework to measure the concentration of an unknown distribution through its i.i.d. samples and a carefully-selected collection of subsets. Our experimental results suggest that the concentration of measure phenomenon is not the sole reason behind vulnerability of the existing classifiers to adversarial examples. In other words, recent impossibility results (Gilmer et al., 2018b; Fawzi et al., 2018; Mahloujifar et al., 2019; Shafahi et al., 2019) should not cause us to lose hope in the possibility of finding more robust classifiers.

Acknowledgements. This work was partially funded by an award from the National Science Foundation SaTC program (Center for Trustworth Machine Learning, #1804603), an NSF CAREER award (CCF-1350939), and support from Baidu, Intel, and Amazon.

References

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.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, 2019.

Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, 2019.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. PAC-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, 2018.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer Science & Business Media, 2013.

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.

David Eisenstat and Dana Angluin. The VC dimension of k-fold union. *Information Processing Letters*, 101(5):181–184, 2007.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. In *Advances in Neural Information Processing Systems*, 2018.

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*, 2018a.

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

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

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

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.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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.

Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist, 2010.

Michel Ledoux. *The Concentration of Measure Phenomenon*. Number 89 in Mathematical Surveys and Monographs. American Mathematical Society, 2001.

Paul Lévy. *Problèmes concrets d'analyse fonctionnelle*, volume 6. Gauthier-Villars Paris, 1951.

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.

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.

Vitali D Milman and Gideon Schechtman. *Asymptotic theory of finite dimensional normed spaces*. Springer-Verlag, 1986.

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.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, 2017.

Stephen M Omohundro. *Five balltree construction algorithms*. International Computer Science Institute Berkeley, 1989.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, October 2011.

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

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

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.

Clayton D Scott and Robert D Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7(Apr):665–704, 2006.

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

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

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

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

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.

Shiqi Wang, Yizheng Chen, Ahmed Abdou, and Suman Jana. MixTrain: Scalable training of formally robust neural networks. *arXiv preprint arXiv:1811.02625*, 2018.

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

Eric Wong, Frank R Schmidt, Jan Hendrik Metzen, and Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, 2019.

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.

A Proofs of Theorems in Section 3

In this section, we prove Theorems 3.3, 3.5, 3.8 and 3.10.

A.1 Proof of Theorem 3.3

Proof. Define $g(\mu, \alpha, \epsilon, \mathcal{G}) = \operatorname{argmin}_{\mathcal{E}\in\mathcal{G}} \{\mu(\mathcal{E}_\epsilon) \colon \mu(\mathcal{E}) \geq \alpha\}$, 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

$$\Pr_{S\leftarrow\mu^m}[\mu(\hat{\mathcal{E}}_\epsilon) \leq h(\mu, \alpha - \delta, \epsilon, \mathcal{G})] \leq \phi(m, \delta). \tag{4}$$

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

$$\Pr_{S\leftarrow\mu^m}\left[\left|\mu(\hat{\mathcal{E}}_\epsilon) - \hat{\mu}_S(\hat{\mathcal{E}}_\epsilon)\right| \geq \delta\right] \leq \phi_\epsilon(m, \delta). \tag{5}$$

Combining Equation 4 and Equation 5, 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

$$\Pr_{S\leftarrow\mu^m}[h(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G}) \leq h(\mu, \alpha - \delta, \epsilon, \mathcal{G}) - \delta] \leq \phi(m, \delta) + \phi_\epsilon(m, \delta). \tag{6}$$

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

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

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,

$$\Pr_{S\leftarrow\mu^m}[\hat{\mu}_S(\mathcal{E}_\epsilon) \leq h(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G})] \leq \phi(m, \delta). \tag{7}$$

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

$$\Pr_{S\leftarrow\mu^m}[|\mu(\mathcal{E}_\epsilon) - \hat{\mu}_S(\mathcal{E}_\epsilon)| \geq \delta] \leq \phi(m, \delta) + \phi_\epsilon(m, \delta). \tag{8}$$

Combining Equations 7 and 8, 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

$$\Pr_{S\leftarrow\mu^m}[h(\mu,\alpha+\delta,\epsilon,\mathcal{G}) \leq h(\hat{\mu}_S,\alpha,\epsilon,\mathcal{G})-\delta] \leq \phi(m,\delta)+\phi_\epsilon(m,\delta). \tag{9}$$

Now combining Equations 6 and 9, by union bound we have

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

□

A.2 Proof of Theorem 3.5

In this section, we prove Theorem 3.5 using ideas similar to ideas used in Scott & Nowak (2006). Before proving the theorem, we lay out the following lemma which will be used in the proof.

Lemma A.1 (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.

Now we are ready to prove Theorem 3.5.

Proof of Theorem 3.5. Define E_T to be the event that

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

Based on Theorem 3.3 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 A.1, 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

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

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

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

□

A.3 Proof of Theorem 3.8

Proof. This theorem follows from our general Theorem 3.5. We show that the choice of parameters here satisfies all four conditions of Theorem 3.5.

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 Eisenstat & Angluin (2007)). 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 3.5 are satisfied. The fourth condition is also satisfied by the universal consistency of histogram rules (See Devroye et al. (2013), Ch. 9). □

A.4 Proof of Theorem 3.10

Proof. Similar to Theorem 3.8 This theorem follows from our general Theorem 3.5. We show that the choice of parameters here satisfies all four conditions of Theorem 3.5.

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 3.8, 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 Eisenstat & Angluin (2007)). Therefore, by VC inequality we have complexity penalties similar to those of Theorem 3.8 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 3.5 are satisfied. The fourth condition is also satisfied by the universal consistency of kernel-based rules (See Devroye et al. (2013) , Ch. 10). □

B The Proposed Algorithms

This section provides the pseudocode and a runtime analysis for our algorithms for finding robust error regions under ℓ_∞ and ℓ_2, respectively.

B.1 Pseudocode

B.2 Runtime Analysis

For ℓ_∞, we construct the systems of hyperrectangles by first precomputing an approximate k-NN distance estimate using Ball Trees (Omohundro, 1989; Pedregosa et al., 2011) for each data point, and then clustering the top-q densest data points into T partitions using the k-means algorithm, where we binary search for the optimal parameter q. The time complexity of precomputing and sorting the nearest neighbor distance estimates is approximately $O(nd\log(n))$, where n is the total number of data points in $\mathbb{R}^d$. In addition, the time complexity of k-means algorithm is $O(ndTI)$, where I is the averaged number of iterations for k-means algorithm to converge. Therefore, the total time complexity of the proposed algorithm for ℓ_∞ is $O(nd\log(n) + ndTI\log(1/\delta))$. In our experiments on CIFAR-10 ($\epsilon_\infty = 8/255$, $T = 40$ and $\delta = 0.005$), the proposed algorithm takes 76 minutes for precomputing the nearest neighbors, and takes around 2 hours for the iterative steps to converge on a Intel Xeon CPU E5-2620 v4 server with 32 processors.

For ℓ_2, instead of computing the k-NN distances for each iteration, we precompute and keep the k-NN neighbours using Ball Trees for each image to save computation, which requires a time complexity of $O(nd\log n)$. The iterative steps require the major computation of $O(\alpha Tn^2d)$, since we iterate through all the possible choices of ball centers and corresponding radii to find the optimal error region with the smallest expansion. We believe the quadratic dependency on the sample size can be

Algorithm 1: 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 : $(\hat{q}, AdvRisk_{\hat{q}}, \mathcal{E}_{\hat{q}})$

Algorithm 2: 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 : $\hat{\mathcal{E}}$

improved using better searching algorithm for finding the robust error region. Since our main focus is to understand the limitation of robust learning on real datasets, we leave the optimization of the proposed heuristic method for better computational efficiency as future work.

C Other Experimental Results

C.1 Results for ℓ_∞ on other datasets

We also evaluate the proposed empirical method for ℓ_∞ metric on other benchmark image datasets, including Fashion-MNIST (Xiao et al., 2017) and SVHN (Netzer et al., 2011).

Table 4: Summary of the main results using our method for different settings with ℓ_∞ perturbations.

Dataset	$\boldsymbol{\alpha}$	$\boldsymbol{\epsilon_\infty}$	$\boldsymbol{T}$	**Best** $\boldsymbol{q}$	**Empirical Risk** (%)		**Empirical AdvRisk** (%)	
					training	testing	training	testing
Fashion-MNIST	0.05	0.1	10	0.758	5.64 ± 0.78	5.92 ± 0.85	10.30 ± 0.72	11.56 ± 0.84
		0.2	10	0.726	5.79 ± 1.00	6.00 ± 1.02	13.44 ± 0.60	14.82 ± 0.71
		0.3	10	0.668	5.90 ± 0.94	6.13 ± 0.93	17.46 ± 0.53	18.87 ± 0.66
SVHN	0.05	0.01	10	0.812	5.21 ± 0.19	8.83 ± 0.30	6.08 ± 0.20	10.17 ± 0.29
		0.02	10	0.773	5.31 ± 0.12	8.86 ± 0.20	7.76 ± 0.12	12.46 ± 0.15
		0.03	10	0.750	5.15 ± 0.13	8.55 ± 0.22	8.88 ± 0.13	13.82 ± 0.25

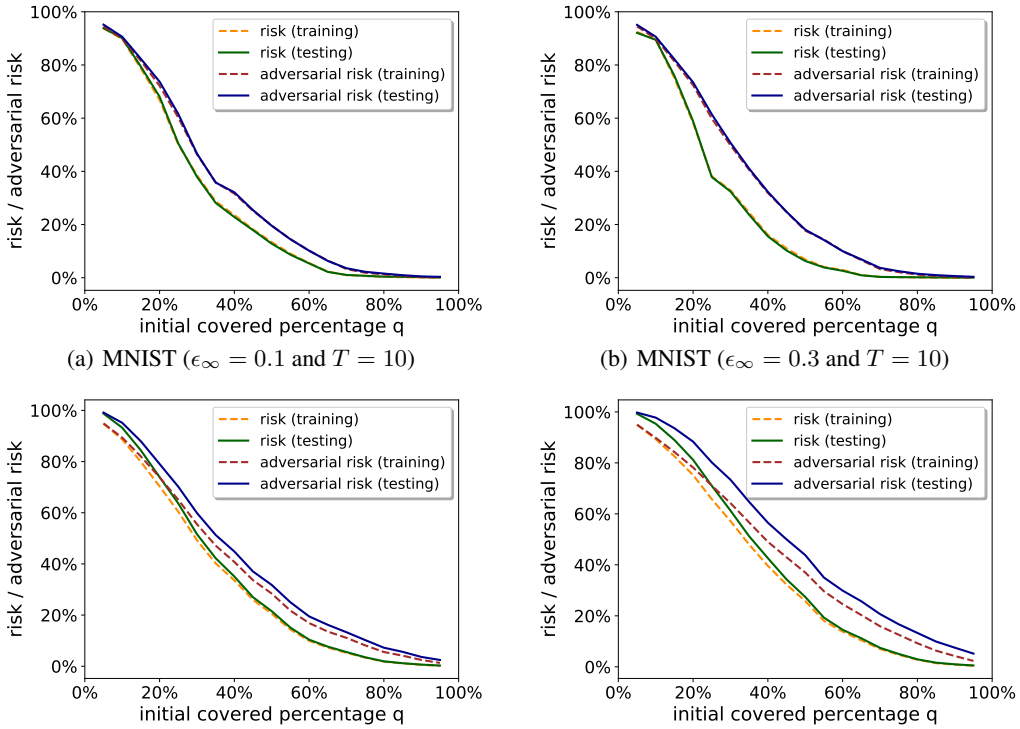
C.2 Detailed results for ℓ_2 using our method

In this section, we demonstrate the detailed training and testing results on the best error region obtained using Algorithm 2 on MNIST and CIFAR-10 with ℓ_2 perturbations, as well as results on Fashion-MNIST and SVHN. Note that for the additional datasets, we set α to be the same as the case of ℓ_∞ and set $\epsilon_2 = \sqrt{n/\pi} \cdot \epsilon_\infty$ using the same conversion rule, where n is the input dimension.

Table 5: Summary of the main results using our method for different settings with ℓ_2 perturbations.

Dataset	$\boldsymbol{\alpha}$	$\boldsymbol{\epsilon_2}$	$\boldsymbol{T}$	**Empirical Risk**		**Empirical AdvRisk**	
				training	testing	training	testing
MNIST	0.01	1.58	20	1.25%	1.07%	2.23%	2.19%
		3.16	20	1.25%	1.02%	4.35%	4.15%
		4.74	20	1.25%	1.07%	10.71%	10.09%
CIFAR-10	0.05	0.2453	5	5.00%	5.16%	5.22%	5.53%
		0.4905	5	5.00%	5.14%	5.61%	5.83%
		0.9810	5	5.00%	5.12%	6.38%	6.56%
Fashion-MNIST	0.05	1.58	10	5.25%	5.07%	7.84%	7.77%
		3.16	10	5.25%	4.99%	15.95%	16.23%
		4.74	10	5.25%	5.21%	19.76%	20.10%
SVHN	0.05	0.3127	10	5.00%	6.92%	5.24%	7.34%
		0.6254	10	5.00%	7.30%	5.59%	8.16%
		0.9381	10	5.00%	7.56%	5.96%	8.94%

C.3 Additional training curves



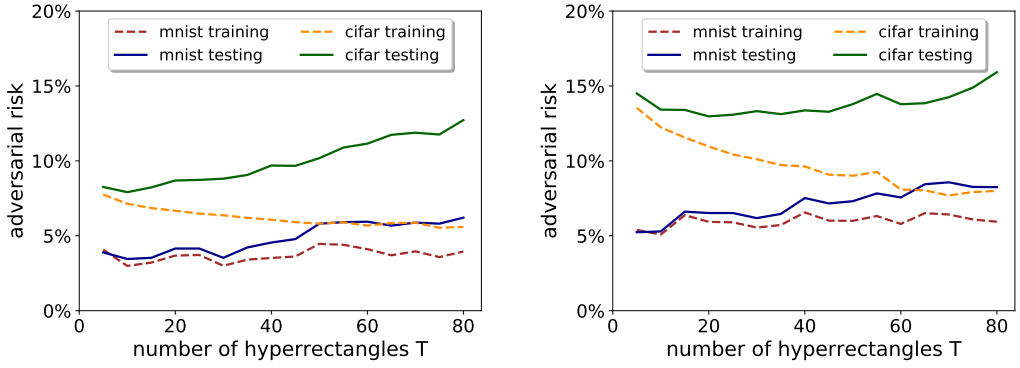
(a) MNIST ($\epsilon_\infty = 0.1$ and $T = 10$)

(b) MNIST ($\epsilon_\infty = 0.3$ and $T = 10$)

(c) CIFAR-10 ($\epsilon_\infty = 4/255$ and $T = 20$)

(d) CIFAR-10 ($\epsilon_\infty = 8/255$ and $T = 40$)

Figure 2: Risk and adversarial risk of the corresponding region as q varies under different settings.



(a)

(b)

Figure 3: Adversarial risk of the resulted error region with best q obtained using our method as T varies under different settings: (a) MNIST ($\epsilon = 0.1$, $\alpha = 0.01$) and CIFAR-10 ($\epsilon_\infty = 2/255$, $\alpha = 0.05$); (b) MNIST ($\epsilon_\infty = 0.2$, $\alpha = 0.01$) and CIFAR-10 ($\epsilon_\infty = 4/255$, $\alpha = 0.05$)